Protein Sequence and Structure Comparison based on Vectorial Representations

Vom Fachbereich Physik
der Technischen Universität Darmstadt

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
genehmigte Dissertation

von
Dipl.-Phys. Florian Teichert
geboren in Frankfurt am Main

Darmstadt 2009
D17

1. Gutachten: Prof. Dr. rer. nat. Markus Porto
2. Gutachten: Prof. Dr. rer. nat. Barbara Drossel

Tag der Einreichung: 2. Dezember 2008
Tag der Prüfung: 16. Februar 2009

Abstract

Proteins are very complex physical objects consisting of thousands of atoms and hundreds of amino acids with complicated local and global interactions on length scales ranging from the microscopic neighbourhood of atoms to the macroscopic size of organisms. The spatial configuration, in spite of that, is encoded into one single character per amino acid using a twenty character alphabet, an apparent contradiction that is not fully understood to date.

This thesis is concerned with problems of protein structure and the relationship of protein sequence and structure. It is tried to integrate the different approaches typically carried out by physicists in the field that investigate very simplified model systems, e.g. single α-helices, with the bioinformatics approach to build powerful analysis tools. The first approach often leads to oversimplified systems that do not describe native proteins as a whole, while the second can be too heuristic and too involved to answer fundamental questions.

We start from defining vectorial descriptions of protein structure, similar in form to sequence descriptions, to firstly compare protein structures, i.e. to perform structure alignments, and discuss several measures for structural similarity. From these we derive a statistical structural similarity score for pairs of protein structure based on their spatial superimposition.

Then we utilize a previously known ansatz to exploit the sequence to structure correlation in order to predict vectorial structure descriptions from protein sequence. These predicted profiles are then used within the same alignment framework to align protein sequences. For these alignments a basic evolutionary similarity measure between protein sequences is derived.

Large part of this thesis is dedicated to the objective assessment of alignment methods including the new method presented and a number of establish programs.

A commonly used measure of structural similarity, the Percentage of Structural Identity (PSI), is discussed and generalized to cover an internal degree of freedom in structure that was ignored formerly. The improvement is achieved by very simple but powerful reasoning. The resulting scheme is also applicable to detect hinges in protein structures.

Concluding, we state that protein structure, despite its complexity, is indeed to a large extent one-dimensional. The unification of structure and sequence alignments under a single formalism gives some insight into the relation of sequence and structure in proteins.

Zusammenfassung

Proteine sind äußerst komplexe physikalische Objekte die aus tausenden von Atomen und hunderten von Aminosäuren zusammengesetzt sind, mit komplizierten lokalen und globalen Wechselwirkungen über alle Längenskalen. Diese reichen von der mikroskopischen Ebene einzelner Atome bis zur makroskopischen Ebene ganzer Organismen. Im Gegensatz dazu kann ihre räumliche Konfiguration in der Sequenz, in nur einem einzigen Buchstaben pro Aminosäure kodiert werden. Dieser scheinbare Widerspruch ist bis heute nicht völlig verstanden.
Diese Arbeit beschäftigt sich mit Fragestellungen aus den Bereichen Proteinstruktur und Protein Struktur/Sequenz Beziehung und unternimmt dabei den Versuch verschiedene Ansätze zu vereinen. Physiker, die in diesem Feld arbeiten, tendieren dazu sehr reduzierte Modellsysteme, wie etwa nur einzelne α-Helices, zu beschreiben, während Bioinformatiker leistungsstarke Analysewerkzeuge entwickeln. Erstere beschreiben häufig so stark vereinfachte Systeme, daß die Ergebnisse nur wenig über reale Proteine aussagen, während letztere oft zu so komplizierten und heuristischen Lösungen kommen, daß keine fundamentalen Fragen mehr beantwortet werden.
Zu Anfang definieren wir vektorielle Proteinstruktur-Darstellungen, in ihrer Form ähnlich zu Sequenzdarstellungen, um in einem ersten Schritt Proteinstrukturen zu vergleichen, d.h. Alignments durchzuführen, wobei auch einige Strukturähnlichkeitsmaße diskutiert werden. Von diesen leiten wir statistische Signifikanzmaße ab, die auf der räumlichen Superposition von Strukturpaaren beruhen.
Im folgenden verwenden wir einen bekannten Ansatz, um aus der Sequenz die vorher definierten Strukturprofile vorherzusagen, die dann mit Hilfe des zuvor für Strukturalignments definierten Algorithmus für Sequenzalignments verwendet werden können. Von diesen Sequenzalignments leiten wir ein Maß für den evolutionären Abstand der betreffenden Sequenzen ab.
Viel Aufmerksamkeit wird der objektiven Beurteilung von Alignment Methoden geschenkt, die Analyse umfaßt dabei den hier vorgestellten Algorithmus und einige bereits etablierte Programme zum Vergleich.
Ein weit verbreitetes Maß für strukturelle Ähnlichkeit, der Prozentsatz struktureller Ähnlichkeit (PSI), wird diskutiert und verallgemeinert um das Auftreten innerer Freiheitsgrade in den Strukturen zu erfassen, die vorher keine Beachtung fanden. Die Verbesserung wird dabei durch einfache aber mächtige Argumentation erreicht. Das resultierende Schema kann auch zur Bestimmung flexibler Drehachsen in Proteinen, sogenannter Hinges, verwendet werden.
Zusammenfassend stellen wir fest, daß Proteinstruktur trotz ihrer Komplexität im Grunde weitgehend eindimensionalen Charakter hat. Die vereinheitlichte Sicht auf Struktur- und Sequenzalignments erlaubt einen Einblick in die Beziehung zwischen Sequenz und Struktur in Proteinen.

Contents

List of Figures

List of Tables

1 Introduction and Motivation – The Biophysics of Proteins

Still today the typical well-defined physical systems with strict equations and closed solutions are restricted to very small numbers of particles and interactions. Already for one of the prototypes of classical systems, the planetary system, only approximate solutions are possible. The other side of the complexity scale is the realm of statistical physics with typically $N_{\mathrm{A}} \sim 10^{23}$ particles. These systems are accessible through strong symmetries and simplifications that lead e.g. to mean-field descriptions.

From the point of view of complexity protein structures are fairly ambiguous objects. On the one hand, protein structures consist of around 20–2000 amino acids, each in turn comprised of 6–15 atoms[1] with various interaction types including peptide bonds, hydrogen bonds, disulphide bonds, coulomb interaction, polar interaction, and salt bridges, making them very complicated complex objects, inaccessible by the strict equations of classical or even quantum mechanics. Amino acids are the building blocks of the protein establishing a stable backbone and defining the protein's sequence by the particular succession of residues. On the other hand, proteins do not show these well-defined symmetries that could be exploited to derived simplified mean-field equations that describe their properties well. Only very involved partly heuristic functions, so-called force-field functions, are known that describe e.g. the free energy of a native folded protein.

The example in Fig. 1.1 demonstrates this complexity. The already rather complicated protein chain PDBid 1floA consisting of many α-helices and two anti-parallel β-sheets is only part of a protein complex of twelve chains. The terms helix and sheet refer to the typical short-range order shapes adopted by the protein chain. While helices are curled like a corkscrew, sheets are folded up and down to the shape of a jigsaw. Together with simple loops, helix and sheet denote the set of secondary structure elements abundantly found in proteins. In contrast to this overwhelming diversity, proteins as well exhibit a high level of order over the whole length scale. Only 20 different amino acid residue types are found at the microscopic scale which

[1] precisely: 6–15 heavy atoms, i.e. other then hydrogen; including hydrogen the numbers rise to 10 atoms for glycine and 27 atoms for tryptophan

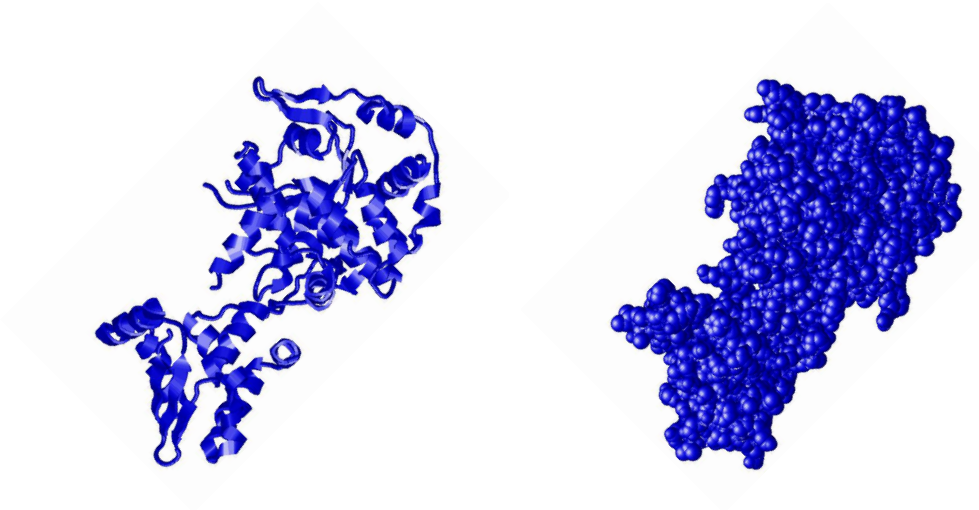


Figure 1.1: The figure shows the molecule PDBid 1floA, $N = 422$. The so-called cartoon display on the left hand side highlights secondary structure elements, while the right picture shows the approximated electron distribution. Pictures are made using the molecule viewing program RASMOL by Sayle & Milner-White [1995].

are connected along the protein's backbone through a stringent sequence of three bonded atoms (i.e. N $\rightarrow$ C_α $\rightarrow$ C). The backbone itself shows strong local order by forming secondary structure elements, α-helices and β-strands, that are in turn folded onto each other to form redundant motifs of tertiary structure. These are then arranged to protein complexes or build into membranes of macroscopic size.
After all, this is a surprising observation: These very complex objects described seem to follow an abundance of 'rules' to accomplish a high degree of order in the folded protein but can be exhaustively described by a simple letter code, i.e. the sequence.
Unfortunately, these rules are not very clear cut. Of course, the collective influence of peptide bonds results in a certain stiffness of the chain that, together with simple volume exclusion, restricts the conformational space accessible by a specific protein chain. Also the network of hydrogen bonds allows to describe the formation of secondary structures and is known to largely stabilize the folded state. However, the process of protein folding cannot be explained from first physical principles, even when ignoring quaternary structure, the formation of intermolecular complexes consisting of distinct protein chains.
Another interesting property of the native folded state is its low binding energy the so-called 'folding free energy'. Typical proteins have folding free energies of about the strength of only one or two hydrogen bonds above body temperature which

means that they are only marginally stable [Lesk, 2002]. Nevertheless, proteins have to be stable to function properly but they also have to have a high 'misfolding energy gap', i.e. a relatively large difference between the state of the lowest folding free energy and higher states, to reliably fold into one and the same structure for a given sequence and avoid massive production of degenerate individuals.
It is clear that a system obeying all this complicated rules at the same time is only accessible through a well chosen set of simplifying models, heuristics, and statistics. As of October 2008 approximately 13 million protein sequences and nearly 55000 protein structures and complexes are experimentally determined. Databases like the PDB [Berman *et al.*, 2000] make all known structures freely available on the Internet and render large scale analyses possible.
In this thesis we formulate and use simplified models to work on different questions of protein science in order to deduce the most information possible with the least data input necessary. Doing so we try to integrate solid physical reasoning, using well-defined models with the bioinformatics demand to create powerful tools that allow for automated statistical analyses. This combines the advantages of both approaches leading to analysis tools from which objective conclusions can be drawn.
Alignments are a crucial tool for many analyses in the fields of biophysics, bioinformatics, and medicine. High quality tools are needed to exploit experimental data to the maximum possible. At the same time also high computation speed is demanded due to the mass of available data. A whole lot of tools are available that may be accurate and fast but their respective abilities are not assessed. In addition, many algorithms are so involved that a potential user is not able to fully understand their functioning. This situation gets even worse when sequence and structure data is being used in the same project. Three different tools are needed then, one for structure, one for sequence alignments, and another for sequence to structure alignments. A thorough selection of these three tools is impossible when respecting possible side-effects induced by a specific combination.
In order to overcome these obstacles we aim to unify the view on alignments motivated by the insight that protein sequence and structure are only two different descriptions of the same physical object, an alignment tool should reflect that analogy. The result of such an ansatz should be a fast high quality algorithm that is able to deal with both structure and sequence data. Aside from that the expert knowledge implemented should be restricted to the minimum necessary to fulfil the requirements in order to keep the scheme understandable.
When investigating alignments from this generalized viewpoint a number of vital topics in current research are being touched: The prediction of the structural profiles gives hints for factual structure prediction projects, one of the most active areas in protein science today. Also the preliminary stage aiming at the prediction of approximate structural descriptions is studied by many groups. Another pivotal topic

is the classification of protein sequences and structures. Databases are growing at high speed assigning the vital task of ordering the mass of available data. A classification is based on a significance score assigned to protein pairs by an alignment algorithm therefore the quality of the classification will be a result of the quality of these alignments.
The 'Structural Classification of Proteins' database SCOP [Murzin *et al.*, 1995] still partly relies on expert knowledge, with all connected drawbacks. A fully automated classification is not available yet.
This and all other mentioned needs are open questions up to some point. Ab inito structure prediction does only work well for limited examples, even though great advances have been made in this field over the past years. Different classification databases exist but keeping them current is not a fully automatic process up to now. Also for the clustering of all known protein structures one has to rely on sequence alignments resulting in much reduced quality only for the sake of computation speed.
This thesis is subdivided into five major parts that coalesce around vectorial descriptions of protein structure and sequence and their application to perform structure and sequence alignments that establish pairwise relationships between amino acids of different proteins. In the first part in Chapter 2 the topology of the protein contact network is described by a number of alternative vectorial profiles. For a limiting case these profiles are nearly identical in their content of information to the structures they are derived from. More involved generalizations allow to describe all globular protein structures retaining the crucial properties of the former definition and showing that these profiles are sound representations of structure. In the second part in Chapter 3 a coarse but highly correlated approximation of these profiles is used to perform protein structure comparisons by aligning structural profiles. To permit the appraisement of the ansatz much effort is made to compare its abilities to a number of well established algorithms, a duty for which no comprehensive reference exists, even though a large number of tools is currently available. The relationship of protein sequence and structure is exploited in the third part in Chapter 4 to predict the structural profile used for structure alignments before from pure sequence information. Doing so also allows to perform sequence alignments with the same algorithm used before. In the fourth part in Chapter 5 an analogue assessment is carried out for the sequence alignments that allows not only to assess and compare the algorithm developed here but also the comparison of the expressiveness of sequence and structure alignments in general. It shows that the sequence alignment introduced leads to even better results then currently available tools from the point of view of detecting structurally relevant similarity in protein sequences.
For both flavours of protein alignment, statistical significance scores are defined that introduce similarity measures for pairs of protein sequences and structures and define a metric quantifying evolutionary and structural distances.

As an addition in the fifth part in Chapter 6 commonly used measures for structural similarity that are based on the spatial superimposition of aligned structures are discussed and improved in a subtle detail, a procedure that also allows to detect conformational changes in protein structure.

2 Vectorial Description of Protein Structure

What is the pivotal property characterizing protein structure? From the first chapter it can be assumed that this question is very hard to address since it does not have a unique answer. For the context of this chapter, however, some properties that a description of protein structure should fulfil are apparent: Unlike e.g. coordinates, a translation/rotation invariant description is demanded due to several reasons. Actually very similar protein structures might differ by torsions of common secondary structure elements or even motifs in respect to each other, besides of the trivial fact that there is no standard definition of a reference frame in which experimental coordinates are stored. When thinking about structural comparisons it is definitely important as well that the description has a certain degree of steadiness against rearrangements in the structure as they happen in protein evolution. Small variations in the structure should lead to small variations in the structural description. This is an essential feature especially for the comparison of only remotely related proteins and renders possible to define a measure to quantify structural similarity in the first place.

The requirement that the description should be as reduced and simple as possible is technically motivated, on the one hand, since this allows for short computation times and renders possible large scale analyses. On the other hand, these reduced descriptions allow more insight in the physical system.

Since the answer to the introductory question, namely the proper choice of protein structure representation, is crucial to the whole project it is thoroughly dealt with in this chapter.

The profiles derived here are lightweight but sound representations of protein structure. They are of vectorial form just like sequence profiles which is a property of major importance for the upcoming sequence alignment task.

2.1 From atomic Coordinates to the Principal Eigenvector

The Protein Data Bank PDB [Berman *et al.*, 2000] is the main source of experimental protein structure data. High resolution coordinate data is available for nearly 55000 protein structures and complexes, mostly analysed by X-ray, NMR, and electron microscopy experiments. The atoms are labelled with atom type (e.g. C, N, O) and functional position (e.g. C_α, C_β) and grouped to amino acids. Amino acid residue type and position in the chain are given which allows to derive the protein sequence.

Proteins are fairly densely packed objects pervaded by a stiff backbone of peptide bonded amino acid monomers. A first reduction in the data is suggested by the strong conformational restrictions imposed by bond lengths, angle limitations, and volume exclusion: Instead of considering all atoms resolved in the experiment, only a single representative coordinate per amino acid can be used. Usually the position of a backbone carbon atom, i.e. C_α or C_β, is used but others are possible, depending on the particular context. This reduction is not totally lossless, even though most of the exact side-chain positions can be recovered with very high precision [Canutescu *et al.*, 2003]. In many contexts, however, this loss can be neglected.

A straightforward and also nearly loss-free simplification[1] of the structural representation that furthermore introduces rotational and translational invariance is achieved by changing from the set of spatial coordinates to the set of pairwise distances yielding the so-called distance matrix (DM).

A distance between amino acids can be defined in many ways and a particular choice is based on the problem to be solved. It turns out that for analyses of evolutionary relationships, a representation based on C_α atoms is more adequate because the side chain configuration changes in the course of evolution. When analysing protein stability, a representation based on all heavy atoms is more suitable.

The next step of reduction is guided by the analogy of protein structure to complex networks. The amino acids mimic the nodes and the pairwise distances play the role of weighted links between these nodes. If only topological properties of the network are in question some of the weakest links can be ignored and weights can be unified without changing the overall network structure.

For the protein (contact) network this transition from the distance matrix to the so-called contact matrix (CM) [Holm & Sander, 1994] is accomplished by applying a θ-function to the components of the distance matrix. All amino acids with mutual distances below a certain threshold distance $d_{\mathrm{th}}^{\mathrm{CM}}$ are said to be *in contact* while those farther away from each other are not. The components of the contact

[1] Chirality is not conserved but can be guessed from the structure in almost all cases.

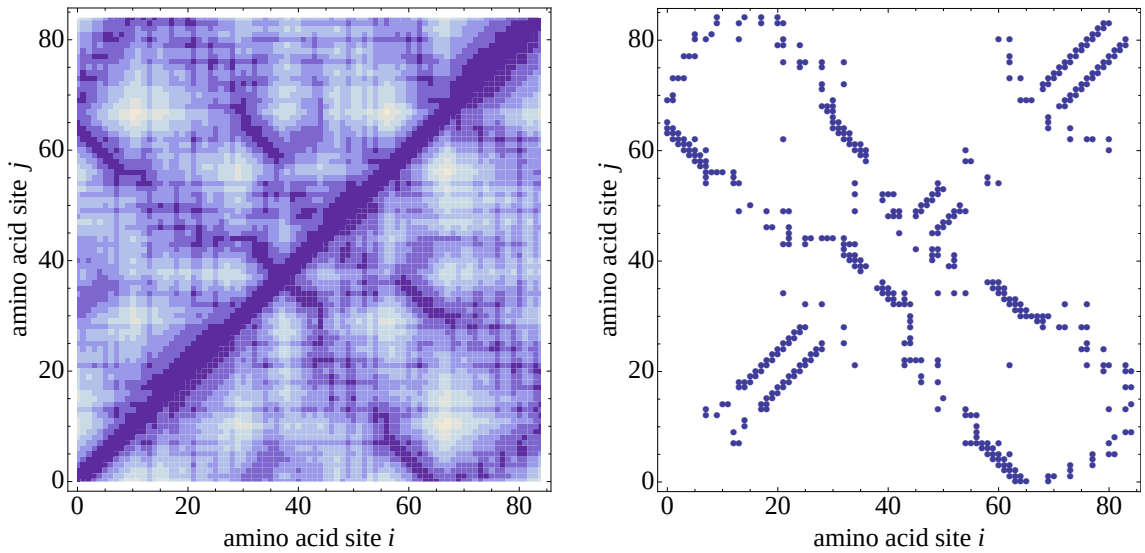


Figure 2.1: The heavy-atoms distance matrix for PDBid 1opdA is shown on the left hand side. The right picture shows the contact matrix derived from it with a distance cut-off of $d_{\text{th}}^{\text{CM}} = 4.5$Å and three excluded diagonals. Dark colour encodes close distances or contacts, while light colours refer to distances up to 32Å or non-contacts. Helices form patterns of continuous contacts directly along the diagonal, while sheets form patterns parallel or orthogonal that can be offset the diagonal.

matrix are consequently set to $C_{ij} = 1$ or $C_{ij} = 0$, respectively. Trivial contacts for amino acids that are already close only due to their proximity in the sequence are suppressed by explicitly setting the main and a number of secondary diagonals n_{D} to zero. This number depends on the distance definition and the threshold applied. That the protein network's topology sufficiently describes protein structure was numerically shown by Vendruscolo *et al.* [1997] where the authors prove that the whole protein backbone can be recovered from the contact matrix[2] with accuracy up to the level of experimental resolution. Figure 2.1 shows such a contact matrix together with the distance matrix it was derived from.

The next step of reduction of the structural representation is motivated by the insight that the main eigenvector of a matrix contains the most prominent features of the matrix itself. The contact matrix C_{ij} is a real (in fact binary) symmetric matrix, it has a real eigensystem. The principal eigenvector, corresponding to the largest eigenvalue λ_1 can without loss of generality be chosen to have components of positive sign only. Each eigenvector component c_i describes the specific amino acid i as row and column i in the contact matrix.

[2]The authors use the term *contact map* which is equivalent to *contact matrix* as used here.

To understand the meaning of these components it is helpful to note that the principal eigenvector c_i maximizes the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$ under the constraint $\sum_i c_i^2 = 1$. For the protein that means that amino acid site i has larger propensity to be close to other sites the larger its value of c_i. Consequently, the term *connectivity* was coined for the principal eigenvector's component c_i that describes the mean contact density amino acid site i feels in its neighbourhood. However, connectivity c_i does not only depend on the local contacts at site i but on the whole arrangement of the structure through maximization criterion Q, making connectivity a global property of protein network topology.
This certainly raises the question whether this reduced profile[3] is really a sufficient description of protein structure. Since that would mean that most of the information of the contact matrix is already contained in its principal eigenvector, ignoring $N - 1$ eigenvectors and N eigenvalues.
The key to answering this question lies in the binary nature of the contact matrix which restricts the set of possible representatives. Although this number is very large, of the order of 2^{N^2}, it was shown by Porto *et al.* [2004] that it is possible to name all contact matrices that match a given principal eigenvector by applying a greedy tree-search algorithm.
The astonishing result of this analysis is that in the overwhelming majority of examples there is only one contact matrix in this set. From the matrix in turn the structure itself can be recovered as mentioned before. Only in some cases, single contacts cannot be determined when the respective amino acids have too little contact to the rest of the structure but these are not structurally relevant anyway.
Albeit several profiles can be derived that associate each protein site i with a single real number v_i it is a unique feature of the principal eigenvector of the contact matrix that its equivalence to protein structure, in the sense used here, can be shown explicitly which encourages to use it as a description of protein structure.
The principal eigenvector (PE) used as a structural representation in this thesis is furthermore normalized to $\langle c_i \rangle = 1$ to make its components independent of chain length.
Another important property of the PE that gets of high importance in the next chapters, is its correlation with protein sequence. For this context we only put on record that the PE allows to predict site-specific probability distributions for a given protein structure, as demonstrated by Bastolla *et al.* [2006]. Figure 2.2 shows the PE of the structure PDBid 1opdA. The associated contact matrix it is computed from is shown in Fig. 2.1.

[3]The term *profile* is used for several descriptions of protein sequence and structure, in this work it is only used for vectorial descriptions.

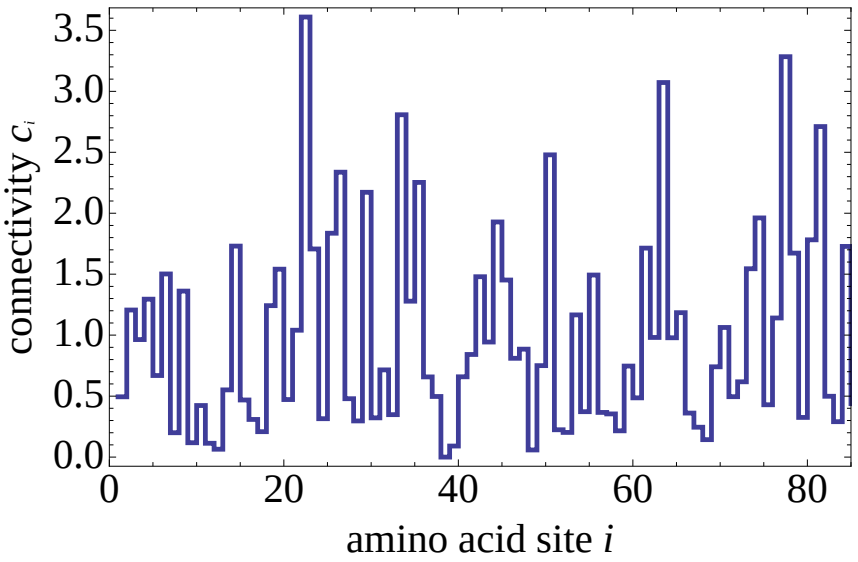


Figure 2.2: The principal eigenvector of PDBid 1opdA is shown, computed from the contact matrix shown in Fig. 2.1. Helices can be recognized by a number of large components interrupted by very small components as in the region 70–84. Sheets are found in regions with consecutive large components as around components 1–10, 30–45, and 60–70.

2.2 Generalization of the PE for multi-domain Structures

The principal eigenvector PE of the contact matrix is a sound representation for protein structure, as seen in the preceding section. Unfortunately, this is only true for smaller single-domain protein structures with little modularity. If structures get larger, compact subgroups form that have significantly higher contact densities within these subgroups than between them. These formations are structural domains or modular substructures below domain level[4].
In the contact matrix representation of multi-domain structures, the matrix splits up into separated modules, one per domain. These modules are only loosely connected to each other. In general, when computing the eigensystem, the principal eigenvector contains only information about the largest, most compact module while information about the remainders are distributed over the whole eigensystem.
In the following, two largely different approaches are described that solve this problem by generalizing the principal eigenvector to multi-domain structures. Although these approaches are of conceptual difference, they lead to highly correlated struc-

[4]to explain why the notion of *domain* is used with some caution, see e.g. Holland *et al.* [2006]

tural representations that approximately reduce to the simple principal eigenvector for small proteins.

2.2.1 The revised Principal Eigenvector

A straightforward ad hoc approach to generalize the principal eigenvector to be applicable to multi-domain structures is to assign a small value $\epsilon(N)$ to those components of the contact matrix whose corresponding sites are not in contact and were formerly set to zero. This small value induces a background connectivity that connects the modules. The error caused by this treatment is small because the numerical value of zero for sites not in contact constituted a strong approximation already in the original definition of the contact matrix. Nevertheless, $\epsilon(N)$ should be selected as small as possible while insuring that all modules get properly connected which equals the requirement that nearly all components of the new vector are non-vanishing.
To fulfil this requirement the length dependent function

$$\epsilon(N) = \min\left\{\epsilon_{\max}, \epsilon_0/\left[\log N - \epsilon_1\right]\right\} \tag{2.1}$$

was fitted over a non-redundant subset of the PDB yielding the values $\epsilon_{\max} = 0.01$, $\epsilon_0 = 0.02$, and $\epsilon_1 = 2$. The revised contact matrix reads

$$\tilde{C}_{ij} = \begin{cases} 0 & \text{for } |i-j| < n_{\mathrm{D}} \\ 1 & \text{for } D_{ij} \leq d_{\mathrm{th}}^{\mathrm{CM}} \\ \epsilon(N) & \text{for } D_{ij} > d_{\mathrm{th}}^{\mathrm{CM}} \end{cases} \tag{2.2}$$

with n_{D} the number of trivial diagonals, D_{ij} the components of the distance matrix and $d_{\mathrm{th}}^{\mathrm{CM}}$ the contact threshold. After computing the principal eigenvector v_i of this revised contact matrix $\tilde{C}_{ij}$, a non-linear transformation $\tilde{c}_i = \mathcal{G}^{-1}(v_i)$ is applied to recover the distribution of the original principal eigenvector's components. This new structural profile $\tilde{c}_i$, that we called the *revised PE* [Teichert & Porto, 2006] (revPE), is nearly identical to the prior principal eigenvector c_i in the limit of small single-domain structures with a typical correlation coefficient of $r = 0.96$ and preserves its predictive power for site-specific amino acid distributions. Hence, the revised definition permits to consistently describe single- and multi-domain protein structures, keeping the crucial properties of the original definition.

2.2.2 The Effective Connectivity Profile

A more systematic alternative to extend the applicability of the principal eigenvector to modular protein structures consists in defining the generalized effective

connectivity profiles (GEC), work that was carried out in collaboration with Ugo Bastolla [Bastolla *et al.*, 2008, 2005]. The GEC profiles constitute a family of vectorial structure profiles whose components c_i self-consistently represent the effective contact density at amino acid site i in native protein structure.
The GEC family of profiles is defined by three rules: All its members share the properties that (a) they maximize the quadratic form $Q = \sum_{ij} C_{ij} c_i c_j$, (b) their mean value is fixed to $\langle c \rangle = 1$ to choose a normalization of the GEC components, and (c) their mean square component is fixed to $\langle c^2 \rangle = B > 1$.
By introducing the two Lagrange multipliers Λ and ϕ that enforce $\langle c^2 \rangle = B$ and $\langle c \rangle = 1$, respectively, these rules can be formulated as

$$\sum_j C_{ij} c_j - \Lambda c_i - \phi = 0 \ \forall i \,, \tag{2.3}$$

with the solution

$$c_i(\Lambda) = \frac{\sum_j (C - \Lambda I)^{-1}_{ij}}{\sum_{kj} (C - \Lambda I)^{-1}_{kj}} \,, \tag{2.4}$$

where I is the identity matrix, M^{-1} represents the inverse of matrix M, and the Lagrange multiplier Λ has to be determined through the constraint on B.
The GEC profiles can be expressed for intuitive understanding as the weighted sum of the eigenvectors of the contact matrix C_{ij},

$$c_i(\Lambda) = \sum_\alpha w_\alpha(\Lambda) \frac{v_i^{(\alpha)}}{\langle v^{(\alpha)} \rangle} \tag{2.5}$$

with the weight coefficients $w_\alpha(\Lambda)$ given by

$$w_\alpha(\Lambda) = \frac{\frac{L \langle v^{(\alpha)} \rangle^2}{\Lambda - \lambda_\alpha}}{\sum_\gamma \frac{L \langle v^{(\gamma)} \rangle^2}{\Lambda - \lambda_\gamma}} \,. \tag{2.6}$$

The weights gradually introduce contributions from higher eigenvectors when the structure described gets more modular. From this one expects that the GEC profiles coincide with the PE for small single-domain structures with low internal modularity if parameter B is chosen accordingly.
For the special choice of $B = B_{\mathrm{cont}} = \frac{\langle \mathrm{cont}_i^2 \rangle}{\langle \mathrm{cont}_i \rangle^2}$ with $\mathrm{cont}_i = \sum_j C_{ij}$, the contact vector, Eq. (2.5) can be solved, explicitly yielding the so-called effective connectivity profile (EC). After computation of C_{ij}'s eigensystem a one parameter optimization routine is used to choose the value of Λ that complies with this choice of B.
The particular form of variance B is motivated by the assumption that structurally exposed amino acids are characterized by being exposed to a lower contact density in comparison to those buried in the structural core. To account for this, parameter B

depends on the surface-to-volume ratio which is approximately given by the contact vector's mean and standard deviation, respectively.
With the alternative definition of $B = B_{\mathrm{PE}} = \left\langle \left(v^{(1)}\right)^2 \right\rangle / \left\langle v^{(1)} \right\rangle^2$, which yields $\Lambda = \lambda_1$, the PE is identified as a member of the GEC family. This analytical connection proves that the PE and the EC profiles are approximately equal for single domain compact structures whose corresponding principal eigenvector has much higher weight than all others.
For the use of the structural profile in the context of structure alignments it is of major importance that it is conserved in evolution. This was explicitly shown by Bastolla *et al.* [2008] for the EC profile by comparing the normalized correlation coefficient of the alignment of c and c' as

$$r(c,c') = \sqrt{N_{\mathrm{a}}} \frac{\sum_i c_{a(i)} c'_{b(i)} - \frac{1}{N_{\mathrm{a}}} \left(\sum_i c'_{a(i)} \sum_i c'_{b(i)} \right)}{\sqrt{\left[\sum_i c^2_{a(i)} - \frac{1}{N_{\mathrm{a}}} \left(\sum_i c_{a(i)} \right)^2 \right] \left[\sum_i c^2_{b(i)} - \frac{1}{N_{\mathrm{a}}} \left(\sum_i c_{b(i)} \right)^2 \right]}} \tag{2.7}$$

with other structural significance scores over a set of structurally related alignments. The sums run over all aligned positions N_{a}, the respective aligned components are indexed by $a(i)$ and $b(i)$. The alignments where computed with the alignment tool MAMMOTH which is independent of the EC profile. Figure 2.3 shows the scatter plot of r over the significance score output by MAMMOTH and the normalized contact overlap score

$$Q(\mathrm{CM}, \mathrm{CM}') = \sqrt{\min\left(N, N'\right)} \cdot q(\mathrm{CM}, \mathrm{CM}') \,, \tag{2.8}$$

over the set of 56450 alignments of structures from the same SCOP superfamilies with less than 40% sequence identity. The strong relatedness of the scores proofs that it is feasible to use the EC itself for structural significance measurement. The correlation coefficients are $r = 0.85$ for MAMMOTH score and contact score, $r = 0.74$ for MAMMOTH score and EC score, and $r = 0.69$ for contact score and EC score.

2.2.3 Revised PE and EC Profile: Two Versions of the same Story

Aside from the PE also the revPE is member of the GEC family of profiles for the special choice of Lagrange multipliers $\phi = L\epsilon\left\langle c \right\rangle (1-\epsilon)$ and $\Lambda = \lambda_1$.
This deep connection can be illustrated by plotting the revPE components as a function of the EC components showing a dependence of sigmoidal shape with different steepness for different examples, as shown in Fig. 2.4. This behaviour most

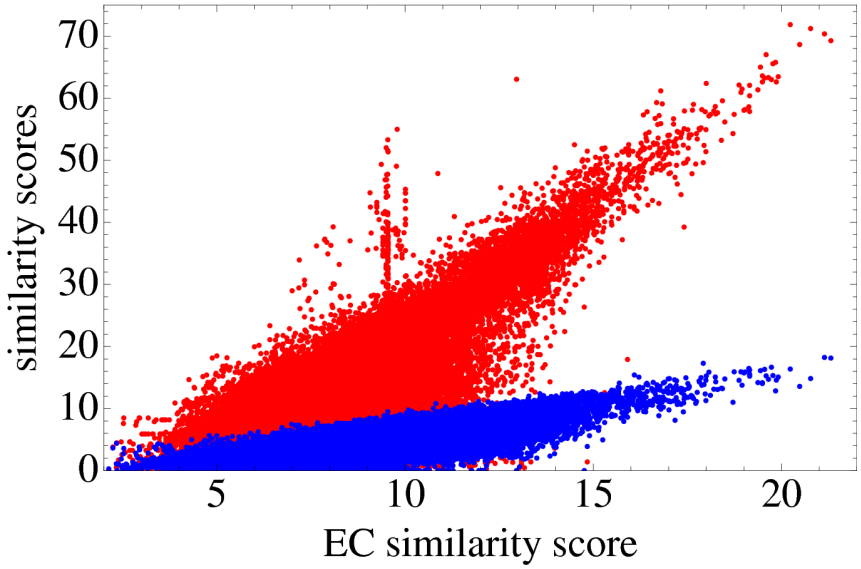


Figure 2.3: Similarity scores between 56450 pairs of evolutionary related proteins in the same SCOP superfamily, with less than 40% sequence identity. MAMMOTH similarity score and contact overlap score, Eq. (2.8), are plotted over EC similarity score, Eq. (2.7), demonstrating that the EC can be used to measure structural similarities. Mutual correlation coefficients are $r(\mathrm{EC}, Q) = 0.69$ and $r(\mathrm{EC}, Z_{\mathrm{MAMMOTH}}) = 0.74$. Data for the plot is taken from Bastolla *et al.* [2008].

likely results from the non-linear transformation used to compute the revised PE. Furthermore, a numerical test reveals very high correlation of $r(\mathrm{EC}, \mathrm{revPE}) = 0.962$ over a non-redundant test set of the PDB.

2.3 The most practical Choice: The Contact Vector

All structural profiles discussed above share one serious drawback for practical applications: They are relatively expensive to compute. For all three a diagonalization of the contact matrix has to be performed, which is a rather lengthy computation even with the most elaborated routines available today. For example, generating the two profiles needed for an alignment takes much longer than running the alignment itself once the profiles are prepared.
A much simpler connectivity related profile is the well known contact vector (CV), already introduced in Section 2.2.2 where it was employed to define the variance B_{cont} that led to the EC profile. The CV is, of course, no member of the GEC fam-

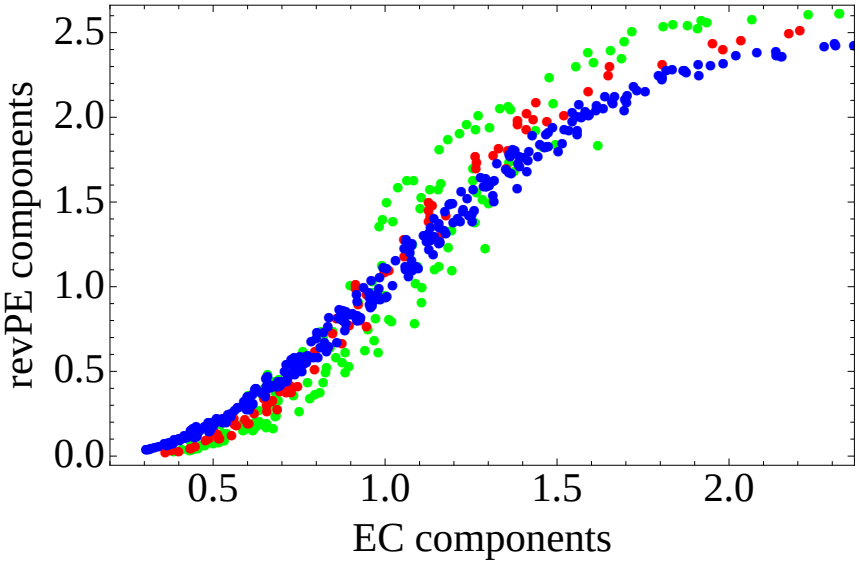


Figure 2.4: The figure shows the revised PE as a function of the EC. The sigmoidal shape is generic for this mapping, while the steepness of the curve depends on the respective example. The picture shows PDBid 1opdA (red), a small single-domain protein with chain length $N = 85$, the myoglobin structure 1a6gA (green) with some internal modularity and $N = 151$, and the larger multi-domain structure 8atcA (blue), chain length $N = 310$. The linear correlation coefficients are $r = 0.965$, $r = 0.951$, and $r = 0.984$, respectively.

ily of profiles, since its components are not real valued but integers simply counting the number of contacts per amino acid site and, hence, do not maximize a quadratic form Q.

Therefore, it is a surprise that the simple contact vector is very highly correlated with the elaborated EC profile (and, consequently, in turn also with the revised PE and the PE for small structures). The correlation of CV and EC is typically as large as $r(\mathrm{CV}, \mathrm{EC}) = 0.933$, as shown in Fig. 2.5. An example for the extent of agreement between EC and CV profiles is shown in Fig. 2.6. Despite the level of agreement between EC and CV profiles it is not expected that the CV is a good candidate for a structural representation. This is due to the CV's inherent degeneracy that follows from its limited set of possible component values. The more so as it is generally not possible to reconstruct the contact matrix from the contact vector. Simply modifying the contact matrix by pairwise exchange of contact indices can leave the contact vector unchanged.

In contrast to that, Kabakçioglu *et al.* [2002] state that the problem of degeneracy is partly compensated by the distinct properties of native protein structure, i.e. the

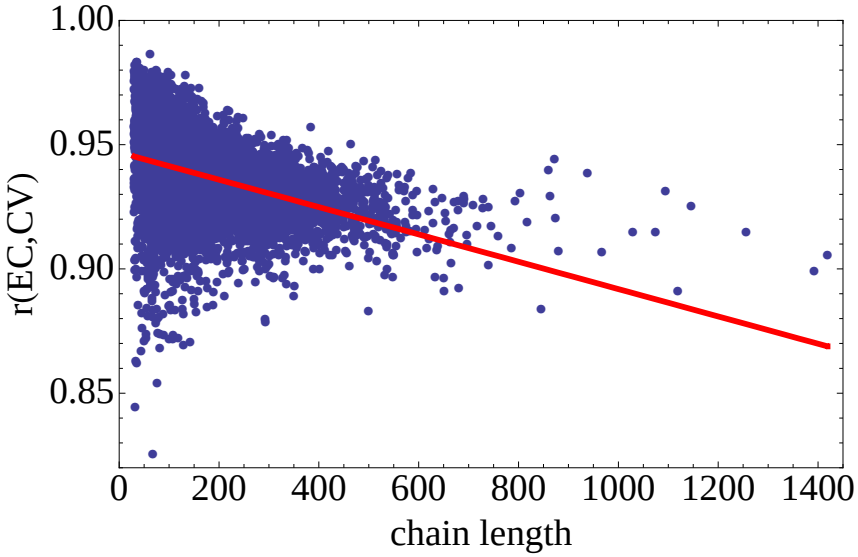


Figure 2.5: The figure shows the scatter plot of the correlation of EC and CV over chain length, for the whole ASTRAL40 database (version 1.73) with 9428 structures. The correlation coefficient $r(\mathrm{EC}, \mathrm{CV}) = 0.933$ in the mean but drops for larger chain lengths.

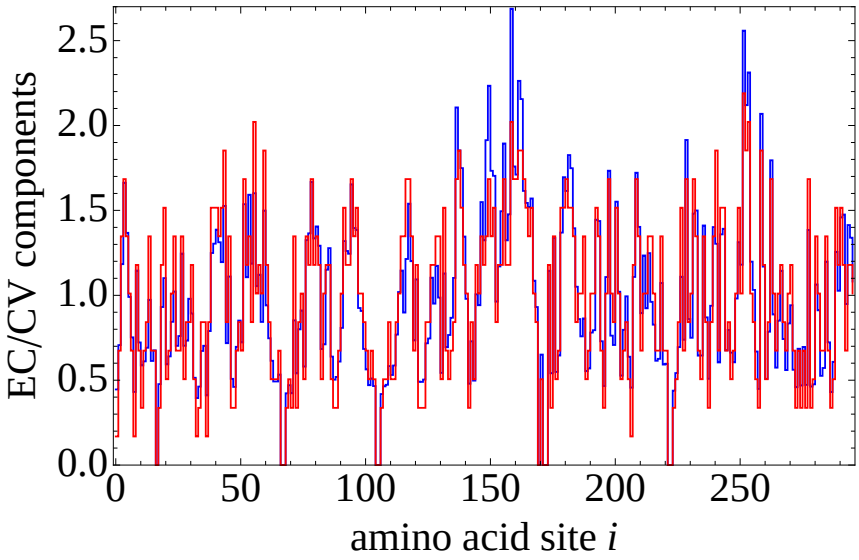


Figure 2.6: EC (red) and CV profile (blue) of the two-domain structure PDBid 1fnbA with chain length 314 amino acids, is shown. Both are based on a heavy-atoms contact matrix with $d_{\mathrm{th}} = 4.5$Å and three suppressed diagonals. The strong correlation between the profiles is obvious.

constraints on the protein backbone rule out most ambiguous cases.
In Chapter 3 we demonstrate in detail that the contact vector is very well behaved in the context of alignments. In fact, the CV leads to very similar results in comparison to the EC profile, suggesting to choose the contact vector as the standard profile for the alignment program developed in the successive chapter.
The particular CV used as a structural representation consists of the number of contacts per site, divided by the mean number of contacts

$$\mathrm{CV}_i = \frac{1}{\langle \mathrm{CV} \rangle} \sum_{j=1}^{N} C_{ij} \,, \tag{2.9}$$

to make the component values independent of chain length. The mean is thereby taken over connected sites i with $\sum_j C_{ij} > 0$, only. Non-connected sites belong to floppy chain parts mostly at the termini of a chain that do not contribute to the stability of the structure and should therefore not affect the component values of the CV.

2.4 Alternative Descriptions of Protein Structure

Many other descriptions of protein structure that do not depend on the notion of connectivity as the ones discussed so far, are defined and commonly used. The distance matrix listing the spatial distances for all pairs of amino acids and different variants of contact matrices were already mentioned in the derivation of the principal eigenvector. The alignment programs DaliLite [Holm & Park, 2000] and TM-align [Zhang & Skolnick, 2005] rely on distance matrices.
The dihedral angles ϕ, ψ, and ω of the protein backbone are frequently used e.g. to visualize the secondary structure content of a protein in a so-called Ramachandran plot. Also relative angles between consecutive C_α-atoms can be used for structure alignments, as done e.g. in the alignment tool MAMMOTH by Ortiz *et al.* [2002].
Furthermore, a lot of tailor-made descriptions exist like the *environmental profiles* used for structure alignments in the alignment program SHEBA [Jung & Lee, 2000] that consist of a mixture of sequence information, secondary structure propensities, and local tertiary structure data.

3 Protein Structure Comparison

The task of comparing protein structures raises three major questions: Firstly, an alignment of two structures has to be computed which means to assign counterparts to the amino acid sites in the two structures or insert a gap if there is none. Secondly, a measure has to be defined that quantifies the similarity of the aligned structures in an objective way and, thirdly, a statistical significance score for structural similarity is required to compare alignments among each other.
The term *alignment* itself can be defined in many different ways. Here, we aim at global pairwise structure alignments without accounting for sequence duplication or exchange. Global, in this context, means trying to find the best match of the full structures as given. This is in contrast to local alignments for which fragments of the structures can be selected, possibly leading to several independently aligned fragments that may be partly redundant or even contradicting. In a global alignment each amino acid site is assigned one or no partner but not more, aligned sites are strictly consecutive in both chains, the sequences are kept intact since copying or exchange of fragments is prohibited. Although the latter modifications occur in the evolution of DNA, and in turn in the protein sequences expressed, it does not constitute a serious limitation to omit them, since protein structures suffering from such mutations are only very rarely stable well-folded objects.
All members of the class of alignments discussed here can be created by inserting an arbitrary number of gaps of arbitrary lengths into both chains, as long as facing gaps are prevented. The result of this procedure can be displayed in the form of sequence alignment strings, even though it was computed from structural data. The strings name the amino acid sequences together with the gaps, as shown in Fig. 3.1. The adjective *pairwise* refers to the fact that only two structures are compared at a time. This is no limitation at all since multiple alignments of more than two structures can be computed from a set of all vs. all pairwise alignments.
Once an alignment is obtained a quantifier for structural similarity is asked for. A number of standard procedures to compute such measures already exists.
For the most commonly used quantities, a structural superimposition is computed from which several indicators can be derived like the spatial RMSD of aligned sites (cRMSD) or the percentage of structural identity (PSI) namely the fraction of aligned sites close in space relating to the length of the shorter chain. Alterna-

```
>PDB:1abaA
--------------------MFKVYGYDSNIHKCGPCDNAKRLLTVKKQPFEFINIMPE
KGVFDDEKIAELLTKLGRDTQIGLTMPQVFAPDGSHIGGFDQLREYFK--------
>PDB:1trsA
MVKQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFLEVDV-
--DDAQDVASEAEVKA-------TPTFQFFK-KGQKVGEFSGANKEKLEATINELV
```

Figure 3.1: The alignment of 1abaA and 1trsA is shown in FASTA format as output by SABERTOOTH. This display is generic for structure and sequence alignments.

tively, omitting the structural superimposition, the contact overlap can be computed that relates the number of agreeing contacts of aligned sites to the total number of contacts.
All these indicators can only be used to quantify the similarity of pairs of structures and all dependent highly on chain length. For classification issues, direct comparison of alignments is desirable which necessitates a statistical significance score mostly independent from system variables. This can be achieved by restraining a measure from its random background.
The alignment scheme developed here is based on the vectorial structure profiles derived in the preceding chapter. After showing that these profiles are sound representations for protein structure that are conserved in the evolution of proteins, we are now able to perform structure alignments by aligning these profiles.
Thereby the task of finding similarities and pointing out differences in two protein structures is translated into recognizing similar connectivity patterns in the corresponding structural profiles which vastly reduces the size of the search space from initially comparing three-dimensional objects to comparing one-dimensional profiles.
In this chapter the machinery to perform such alignments is developed, similarity scores are defined and a significance score is deduced, as published by Teichert *et al.* [2007]. Subsequently, the new algorithm's abilities to recognize structural similarity and classify structures are matched with established algorithms for benchmarking.
Aside from judging the quality achieved with the technique developed here, assessment and comparison are important sources of information that provide insight in the abilities currently achieved in the field. To approximately appraise this quality in relation to the best possible quality, best-of reference sets are compiled gathering the best alignments of a number of reference tools.
Although many alignment tools exist today a comprehensive benchmark is not available making the selection of a favourite tool a matter of taste.

3.1 Computing the Alignment

The combinatorial number of possible alignments explodes with the length of the protein chains in the alignment as soon as gaps of some plausible definition are introduced. Therefore, the notion of *alignment* has to be defined pin sharp rendering it algorithmically accessible. From this set of possible alignments, that is still of staggering cardinal number, the one that corresponds to the best similarity match has to be chosen. To achieve an acceptable mapping from the codomain of a cost function to the similarity of actual native protein structures is the pivotal question of this task.

3.1.1 Enumerating all possible Alignments

The alignment of structural profiles can be conducted very similarly to traditional sequence alignments. Every possible alignment that can be constructed by just inserting an arbitrary number of gaps of arbitrary lengths, can be represented by a path through an alignment matrix A_{ij}, as the one depicted in Fig. 3.2. Building up this path, a diagonal step $A_{i-1,j-1} \rightarrow A_{ij}$ means to align amino acid $A_i^{(1)}$ from chain one with $A_j^{(2)}$ from chain two. Horizontal and vertical steps introduce gaps in chain one and two, respectively. The set of admissible paths consists of all possible combinations of consecutive steps starting in the upper left corner of the matrix, ending at the lower right.
The matrix A_{ij} has dimensions $(n+1, m+1)$ with profile length n for protein structure one and m for structure two. The additional first row and column are needed to allow for leading gaps, while trailing gaps are implemented by permitting direct jumps from all elements A_{ij} for which $i = n$ or $j = m$ holds to the End element.

3.1.2 Defining the optimal Profile Alignment

The optimum alignment path minimizes a cost function, which depends on a set of parameters that are analogous to traditional substitution probabilities for alignments and open/extend penalties for gaps as known from sequence alignments. However, in contrast to those, the penalties used here can be directly connected to the structures through their explicit dependence on the profiles' components.
Evidently, the cost function must be such that the alignment of amino acids $A_i^{(1)}$ and $A_j^{(2)}$ is favoured if the associated profile components $c_i^{(1)}$ and $c_j^{(2)}$ are similar, the cost should increase when aligned components get more different.

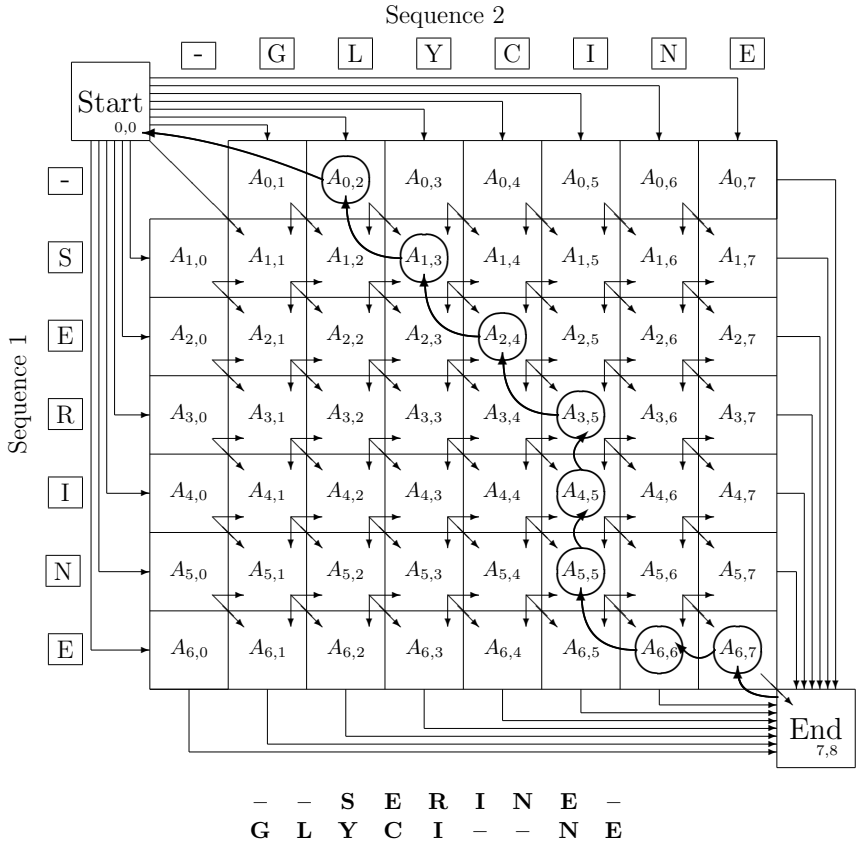


Figure 3.2: The alignment matrix with a possible path encoding a specific alignment is depicted. Diagonal steps correspond to aligned amino acids while horizontal and vertical steps introduce gaps in chain one and two, respectively. The path following arrows from *End* to *Start* refers to the cheapest path found. Below the figure the alignment related to that path is displayed in sequence alignment description.

Inserting a gap is penalized in two different ways. First, the chain in which the gap is inserted needs to be broken. From a structural point of view this is equivalent to disrupting a number of contacts, which is less likely in parts of the chain that are more highly connected, simply because more contacts have to be broken. A second penalty models that it is less likely that the inserted chain part (that is opposite the gap) is very highly connected to the rest of the structure because a higher number of contacts imposes stronger steric constraints. By choosing higher costs for opening a gap than for extending a gap once it is established, a bunching effect is realized that reduces fragmentation of the alignment. In order to account for the fact that proteins are less stable at their termini different weighting and scaling parameters should be chosen inside the chains and at the ends.

Of course, dictions like 'breaking a chain' are a bit sketchy and should be understood from an evolutionary point of view. In this context to break a chain and insert or delete amino acid sites means that corresponding codons were inserted or deleted in the DNA coding for that protein, changing in turn the protein sequence that is expressed. Whether this modified chain is able to fold properly is more likely the less impact the abovementioned modifications have on its structural properties. What is modelled by the set of penalties is this otherwise fairly inaccessible mapping from the evolutionary change in DNA to the change in the folded structure.
The entangled use of these break and insert contributions to the gap penalty reflects the inherent ignorance of whether a gap in the alignment was caused by the deletion of a fragment from one chain, or by the insertion of a fragment into the other chain in the course of evolution. To address this question at least heuristically, a multiple alignment of similar structures would be needed to bring about a majority decision. The penalties that build up the path cost function in detail are divided into four terms:

1. Aligned components of the structural profiles, corresponding to position i in the first profile and position j in the second profile, are penalized by a term that grows with their absolute difference,
$$M_{ij} = \left| c_i^{(1)} - c_j^{(2)} \right|^{p_{\mathrm{align_e}}} \tag{3.1}$$
with $p_{\mathrm{align_e}}$ as a tunable parameter used to scale the contribution.

2. Breaking chain s between residue i and $i+1$ is penalized by
$$B_i^{(s)} = p_{\mathrm{break_f}} \cdot \left(\frac{c_i^{(s)} + c_{i+1}^{(s)}}{2} \right)^{p_{\mathrm{break_e}}} \tag{3.2}$$
with parameters $p_{\mathrm{break_f}}$ and $p_{\mathrm{break_e}}$ and with $s \in \{1, 2\}$ selecting the chain. This is based on the expectation that it is less likely to break a chain at a strongly constrained position with large components c_i and c_{i+1}.

3. An insertion of length n_j in chain s at position $j+1$ opposite to a gap in the other chain, consisting of the components $[c_{j+1}^{(s)} \ldots c_{j+n}^{(s)}]$ is penalized by
$$I_j^{(s)} = p_{\mathrm{insert_f}} \cdot \sum_{k=j+1}^{j+n} c_k^{(s)^{p_{\mathrm{insert_e}}}} \tag{3.3}$$
with parameters $p_{\mathrm{insert_f}}$ and $p_{\mathrm{insert_e}}$ and with $s \in \{1, 2\}$ selecting the chain. This is based on the expectation that strongly connected residues are less likely to form part of an insertion, or to be deleted from the other chain.

4. An additional bit of information can be drawn from the sequence of amino acids itself. Substitution of amino acids is less likely for pairs with very different physiochemical properties just like used in sequence alignments,

$$S_{ij} = p_{\mathrm{AAsubst_f}} \cdot \left(1 - P(A_i^{(1)}, A_j^{(2)})\right)^{p_{\mathrm{AAsubst_e}}} \tag{3.4}$$

   with the parameters $p_{\mathrm{AAsubst_f}}$ and $p_{\mathrm{AAsubst_e}}$ and with $P(A_i^{(1)}, A_j^{(2)})$ for the substitution probability connecting amino acids $A_i^{(1)}$ and $A_j^{(2)}$.

To include a sequence dependent part into the structure alignment looks a bit problematic on first sight and, indeed, this additional information needs to be handled with some care. The major issue is that sequence similarities strongly depend on the evolutionary distance of the whole sequences, not only on the amino acid pairs. Furthermore, related structures can have sequences without measurable sequence similarity leading to a noisy if not wrong signal. Bearing these aspects in mind the influence of sequence information should be kept under some control. As a result, the complete path cost function F is given by

$$\begin{aligned} F = \quad & \sum_{\forall (i,j) \in \mathcal{P}} M_{ij} + \\ & \sum_{\forall i \in \mathcal{B}^{(1)}} B_i^{(1)} + \sum_{\forall i \in \mathcal{B}^{(2)}} B_i^{(2)} + \\ & \sum_{\forall j \in \mathcal{I}^{(1)}} I_{j,n_j}^{(1)} + \sum_{\forall j \in \mathcal{I}^{(2)}} I_{j,n_j}^{(2)} + \\ & \sum_{\forall (i,j) \in \mathcal{P}} S_{ij} \end{aligned} \tag{3.5}$$

where $\mathcal{P}$ is the set of all aligned pairs of amino acids, $\mathcal{B}^{(s)}$ the set of all positions i after which chain s is broken, and $\mathcal{I}^{(s)}$ the set of all insertions of length n after position j in chain s.

3.1.3 The Alignment Algorithm

Being equipped with the playing field and the cost function to define the set of possible outcomes and order them by relevance, the leftover task is to pick out the path with the lowest overall cost that corresponds to the optimum profile alignment per construction, given properly adjusted parameters.
This task can be performed by very efficient standard algorithms like Dijkstra's shortest path algorithm [Dijkstra, 1959], Needleman & Wunsch [1970], or Dynamic Programming, to name only a few.
The resulting alignment can be represented in the form of superposed profiles, as in

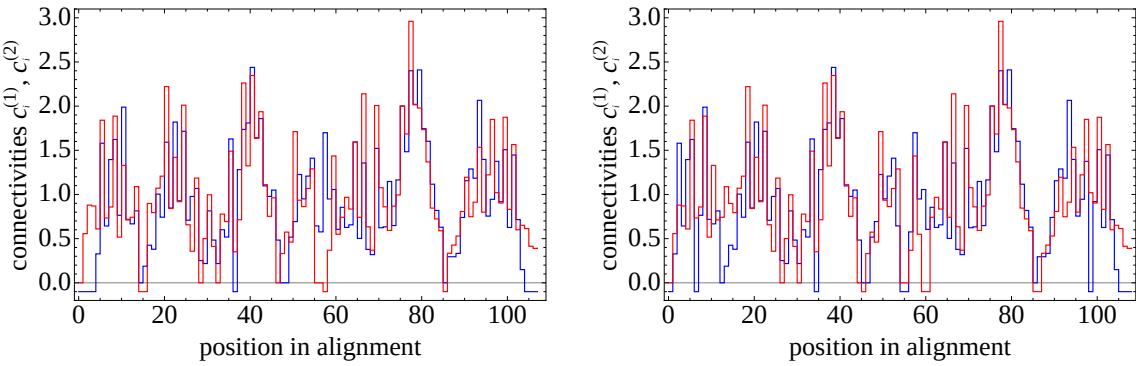


Figure 3.3: The figure on the left shows the optimal profile alignment of the domains ASTRALid d1cd9b2 and d1bpv__ as computed by the profile alignment step of SABERTOOTH. Gaps, as marked by the negative values, are inserted in regions of low connectivity, continuous similar patterns of larger connectivity are correctly aligned. The right figure shows the final alignment after post-processing. The number of residues close in space after optimal rotation increases from 56 to 78 residues. The three-dimensional superimposition of this alignment is shown in Fig. 3.5 on page 33.

Fig. 3.3. As expected, continuous regions of similar patterns are aligned and gaps are inserted in regions of low connectivity. Up to this point the only information used for the alignment is contained in the structural representation, the scoring function with parameters, and the amino acid sequence.

3.1.4 The Parameter Training Scheme

The choice of adequate parameters is a subtle point in the optimization problem presented above. Initially, the free parameters in cost function F in Eq. (3.5) define an 11-dimensional non-linear search space, ruling out any analytic approach. Moreover, the definition of a scoring function and the optimization of its parameters is not straightforward, simply because the best score should correspond to the best structure alignment. For this mapping a good approximation is needed beforehand. In Section 3.2 it will be argued that the PSI is such an approximation, so that the length weighted mean PSI over a training set of alignments seems a good target for parameter training.
Unfortunately, this score is step-valued so that its landscape is expected to be very rough. In addition, large parts of the alignment may jump when changing parameters as it is the result of a global optimization procedure. This effect is even aggravated

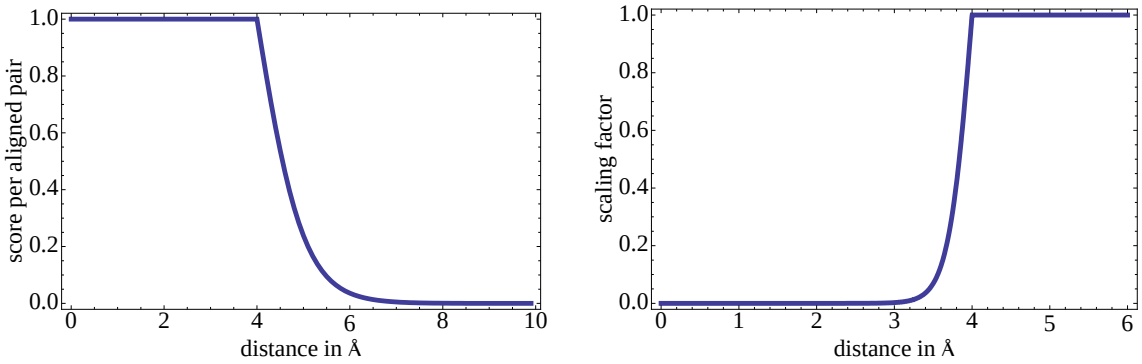


Figure 3.4: The left plot shows the per site contribution to the function optimized in the parameter training scheme. On the right, the scaling function used to modify the cost of allowed alignments in the post-processing step is depicted. Both implement the definition of close contacts $d_{\text{th}}^{\text{PSI}} = 4\text{Å}$ borrowed from the PSI.

by the specific properties of the cost function that force to some decidedness through the bunching tendency imposed by the gap penalties, as discussed above.
To smoothen the scoring landscape and to procure an attractive effect towards better solutions a soft tail is added to the scoring function,

$$O(D_{i;j}|\mathcal{P}_{\{i\}}; R_{\{i\}}, \vec{t}_{\{i\}}; \{\text{p}\}) = \begin{cases} 1 & \text{for } D_{i;j} \leq d_{\text{th}}^{\text{PSI}} \\ 2/(\exp\left\{\frac{1}{2}(D_{i;j} - d_{\text{th}}^{\text{PSI}})\right\} + 1) & \text{for } D_{i;j} > d_{\text{th}}^{\text{PSI}} \end{cases} \quad (3.6)$$

which itself depends on the set of inter-structure aligned distances $D_{i;j}$. These, in turn, depend on the set of aligned residues $\mathcal{P}_{\{i\}}$ and the spatial superimposition specified by rotation matrices $R_{\{i\}}$ and translation vectors $\vec{t}_{\{i\}}$ which furthermore depend on the set of parameters $\{p\}$. Index i runs over all alignments in the training set, while j enumerates all aligned pairs in $\mathcal{P}$. The distance cut-off $d_{\text{th}}^{\text{PSI}} = 4\text{Å}$ is borrowed from the standard definition of the PSI. The soft tail of the scoring function is depicted in Fig. 3.4, its parameters are manually optimized.

3.1.5 Alignment Post-Processing

The profile alignment establishes a relationship between structures by aligning pairs of amino acids. The quality of this alignment can be assessed from the superimposition of the structures in space. How this superimposition is actually created is a subtle point which is discussed in detail later in this chapter. For the application here it is important to note that the relations assigned in the global connectivity

match can be used to derive a spatial rotation in order to obtain detailed local information, namely distances between all amino acid sites in the two structures.
Obviously, if the profile alignment is wrong this first superimposition does not give clear evidence of the structures' similarity. If it is right, grasping the big picture correctly, a full inter-structure distance matrix $D_{i;j}$ directly reveals agreeing parts, common in both structures. This information can be exploited in a second alignment step in order to refine the alignment itself, an idea implemented in many alignment algorithms.
This two-step approach merges the two different points of view on protein structure in an elegant way: The coarse-grained global connectivity fingerprint, embodied in the connectivity profile, points out global conformational agreement, while the precise distance information $D_{i;j}$ describes structural similarities on the level of local details.
This procedure is finally carried out by identifying pairs of C_α atoms that are closer in space than a threshold $d_{\text{th}}^{\text{refine}}$, disregarding whether these pairs were formerly aligned or not. Thus, the profile alignment enters the scheme only through determining the spatial superimposition.
Spurious pairs that might spoil the gain of post-processing can be sorted out by imposing the condition that only pairs that are member of sufficiently long continuous fragments of pairs $l \geq l_{\min}$ are to be regarded. The values of these parameters must be chosen carefully since the result of the whole procedure is strongly influenced by their selection. More and more incidental pairs are selected when increasing the distance threshold and decreasing the minimum fragment length parameter, which would lessen the significance of the alignment. The remaining set of relevant amino acid pairs marks a path through alignment matrix A_{ij}. To force the optimization routine to cross these elements the possible paths simply need to be restricted to that set by disallowing all steps that are not needed to connect allowed segments. In the case of ambiguities, if a site is member of more than one allowed path, the better one can be favoured by modifying the alignment matrix using the distance matrix above $A_{ij} \xrightarrow{D_{i;j}} A'_{ij}$, before the second run of the optimization algorithm. This second run on the restricted and reweighed alignment matrix A'_{ij} defines the final alignment. The same cost function parameters are used for both the initial profile alignment and for the post-processing step.
When applying this post-processing procedure one has to be aware of its limitations. The larger the components $D_{i;j}$ get the less meaningful are they. In fact, very large distances are very likely the result of the fact that the rotation applied to superimpose the structures is not suitable for this part of the alignment. This may be caused by internal movements in the structures rather than real dissimilarities, as discussed in more detail in Chapter 6. In order to keep the scheme general in the

presence of such phenomena a cut-off should be introduced that ensures that too large distances are not taken into account.
It shows that only slight corrections are introduced by the post-processing routine when comparing the results before and after post-processing. In general, the overall similarity match is retained while the improvement is achieved by introducing obvious deviations from the overlap maximization of the structural profiles, as shown in Fig. 3.3. This implies that some local properties of the structures are not represented in the profiles, as expected by their construction.
The whole post-processing procedure relies on the quality of the first profile alignment from which the spatial superimposition is created. In this sense, the problem is solved in the profile alignment step while the post-processing is only needed to improve the local details.

3.1.6 Details of the SABERTOOTH Implementation

Since the alignment framework is mostly generic with respect to the choice of structural profile, a large number of connectivity based profiles could be assessed to take the best choice of parameters in matters of accuracy and speed. Variations of the EC based on C_α and heavy-atom distances, derived from binary and real-valued contact matrices were probed, as well as the revPE and contact vectors based on different contact matrix definitions. Several candidates lead to very similar alignment quality, even though for single alignment examples the difference can be dramatic these variations equal out over the test sets.
After all the contact vector CV on C_α distances with a distance cut-off $d_{\text{th}} = 17$Å and $n_{\text{D}} = 3$ excluded diagonals was selected as the standard structural representation [Teichert *et al.*, 2008], mostly motivated by practical reasons: The contact vector is very efficient to compute and the C_α trace is available for all structures contained in the PDB. Also when the alignment of computationally created decoys, e.g. in a fold prediction experiment, is asked for usually only the backbone is known. The comparatively large distance cut-off for contacts is the result of extensive parameter scanning with different values for the cut-off and the number of excluded diagonals. Although the contact threshold for C_α distances is usually in the area of 8–12Å, it seems reasonable to use a somewhat larger value for this application. Firstly, because the higher number of contacts has a smoothing effect on the profile and, secondly, because in this way a larger neighbourhood of the sites is included putting higher weight on tertiary structure conformation. When scanning parameters for different cut-offs a clear maximum favoured this choice.
The parameters used by SABERTOOTH were trained on a set consisting of 235 superfamily related alignments that were randomly selected from the manually curated

set of 29 representative SCOP superfamilies defined by Leo-Macias *et al.* [2005]. It turned out that the scaling exponents are valid within the structures and at the termini for both insert and break penalties. Different values were found for the weighting factors for chain insertion, as expected from sequence alignment parameter values, even though the difference of the parameters is much smaller than expected from the analogous values there. An additional parameter for weighting the break of the chain at a terminus, i.e. to assign gaps before or after the chain, can trivially be omitted, leaving eight free parameters.
For the actual training an elaborated Monte Carlo scheme with Metropolis criterion and simulated annealing was implemented to deal with the rough landscape of the optimized scoring function.
The computation of the best alignment itself is carried out by a Dijkstra's shortest path routine which allows for the very efficient implementation of the post-processing step. Components of the alignment matrix A'_{ij} that are not allowed anymore after analysing the first spatial superimposition can simply be omitted by deleting their respective connections. Besides of that each optimization algorithm that is able to deterministically find the least cost path would be applicable.
The sets of parameter values used in the cost function, Eq. (3.5), are listed in Table 3.1 for both, the CV, actually used by SABERTOOTH, and for the EC, as an example for a very differently defined profile that is used later on to demonstrate the algorithm's stability against changing the structural representation. The EC is based on a heavy-atoms contact matrix with a contact cut-off of $d_{\mathrm{th}} = 4.5$Å and three excluded diagonals, i.e. it is different in profile and distance definitions, as well as used parameters. The substitution probabilities $P(A_i^{(1)}, A_j^{(2)})$ used in the sequence dependent term S_{ij} are the same used to define the BLOSUM matrices. To recover the actual probability values from the matrix components of freely available BLOSUM matrices is a rather complicated computation procedure that can be carried out by the program *lambda* by Eddy [2004]. Here BLOSUM62 was used, the substitution matrix computed from the BLOCKS database clustered with 62% sequence identity cut-off. Without further knowledge about the evolutionary distances in the alignments this matrix can be considered an all-purpose compromise. Most sequence alignment tools apply this matrix by default.
For the post-processing step some more parameters enter but these are adjustable mainly by thorough reasoning. Close amino acid sites in the alignment should obey the same cut-off as defined for the PSI, i.e. $d_{\mathrm{th}}^{\mathrm{refine}} = d_{\mathrm{th}}^{\mathrm{PSI}} = 4$Å. A consecutive group of close sites should be at least four amino acids long, $l_{\mathrm{min}} = 4$, to suppress spurious contacts that might e.g. be caused by orthogonally crossing helices.
All A_{ij} components that are allowed to be traversed in the post-processing step are defined by these parameters. Furthermore, the alignment cost is reduced before the post-processing run for all sites that are allowed and even closer in space than the

Parameter	$CV_{C_\alpha 17/3}$	$EC_{HA4.5/3}$
p_{align_e}	1.20648	1.42052
p_{break_f}	1.09642	1.91132
p_{break_e} $p_{break@term_e}$	1.60294	1.34184
p_{insert_f}	0.502488	0.660006
$p_{insert@term_f}$	0.335979	0.487867
p_{insert_e} $p_{insert@term_e}$	2.28409	1.63232
$p_{AAsubst_f}$	0.594701	0.4042
$p_{AAsubst_e}$	11.1109	10.035

Table 3.1: The table shows the parameter values for the alignment penalties for breaking a protein chain and inserting a chain fragment opposite to a gap. The same parameters are used in the profile alignment as well as in the post-processing routine. The parameter values are relative to alignments using the contact vector CV with $d_{th} = 17$Å, $n_D = 3$, and the EC on heavy-atoms with $d_{th} = 4.5$Å, $n_D = 3$, respectively.

threshold d_{th}^{refine}, larger distances stay untouched which leads to the transformation

$$A'_{ij} = \begin{cases} \dfrac{\text{scale}}{1 + \exp\{(d_{th}^{refine} - D_{i;j})/\Delta\}} & \text{for } D_{i;j} \leq d_{th} \\ A_{ij} & \text{for } D_{i;j} > d_{th} \end{cases} \tag{3.7}$$

with scale = 2 and $\Delta = 0.15$Å. The steep drop induced by the selected value of Δ, which effectively sets all $A'_{ij} \approx 0$ for $d \lesssim 3$Å as shown in Fig. 3.4, is justified by the cautious selection of allowed alignments. The algorithm is implemented in the software package SABERTOOTH that can be accessed at the group's web server at `http://www.fkp.tu-darmstadt.de/sabertooth/`. All computation shown in this thesis was carried out using it.

3.2 Assessment of Structure Alignments

Evaluating a given structure alignment is not at all a trivial task. Which given alignment is better than another depends on the context and is subjective up to

some point. Because of this no database of objectively optimal alignment examples exists that could be used as a gold-standard.[1] For structure alignments mostly objective measures can be derived from the spatial superimposition of the structures. Nevertheless, *objective* can at most be used in the sense of *commonly used*, since there are ambiguities left that have to be respected.
The RMSD on superimposed structures (cRMSD) is applicable as a distance measure between two structures, with the drawback that it can be trivially minimized by reducing the number of aligned amino acid sites. The PSI, counting aligned sites that are close in space, depends on the choice of distance definition and cut-off and can be artificially increased by increasing the number of aligned sites and by increased fragmentation in the alignment without improving the actual alignment quality, as documented in Teichert *et al.* [2007].
Both measures furthermore depend on the structural superimposition that may itself be computed under different premises, e.g. to either minimize the total cRMSD or maximize the number of close pairs in space, possibly leading to quite different results.
A latent ambiguity of the PSI and alike can be compensated by adopting a sequence dependent measure even on a structure alignment. The commonly used PSI cut-off of $d_{\text{th}}^{\text{PSI}} = 4\text{Å}$ is not sensitive for every case of local shifts by a single position along the protein chain due to the typical C_α–C_α distance of 3.7Å, here sequence similarity can give evidence on which position is more appropriate.
A whole zoo of alternative measures is defined. They are mostly deduced from alternative structural representations like, e.g. the contact overlap that is computed on the contact matrices of aligned structures, counting agreeing contacts normalized by the smaller number of total contacts in the two matrices. All these measures must be considered less fundamental since they strongly depend on the detailed choice of representation. For the contact matrix example the kind of distances used, the contact cut-off, and the number of zeroed diagonals is negotiable, as discussed in Section 2.1. However, the contact overlap has the advantage that it is rotation/translation independent and, hence, suffers less from structural distortions that would reduce the common core identified through spatial superimposition.
Eventually, there is no unique measure evaluating alignment quality and one is left considering a set of different measures. Higher PSI values (better rPSI, see next section), for example, are strictly only better than lower ones if cRMSD and seqSim values are of at least comparable size. Only gradual improvement of this situation is achieved by introducing the so-called Global Distance Test/Total Score (GDT_TS) by Zemla *et al.* [1999], used in the CASP project [Tramontano, 2007].

[1] For sequence alignments the BAliBASE database of manually curated multiple sequence alignments exists but the same arguments apply here as well.

The GDT_TS is a weighted mean of PSI values for different cut-offs, usually defined as GDT_TS $= \frac{1}{4}(\mathrm{PSI}(1\text{Å}) + \mathrm{PSI}(2\text{Å}) + \mathrm{PSI}(4\text{Å}) + \mathrm{PSI}(8\text{Å}))$. One Ångström is below the resolution of most experimental samples while eight Ångström includes sites so distant that their relevance is rather unsure.
All mentioned measures are applicable to pairwise comparison only, since their values depend on system variables like chain length or number of contacts they are not comparable amongst themselves. A measure of significance mostly independent of system parameters is needed for several tasks, e.g. cluster a protein database by structural similarity, the first step to establish a classification. This is a vital issue needed for several partially unsolved tasks in protein science like defining non-redundant subsets of a database, or identifying the 'building blocks' of proteins, i.e. common motifs redundant throughout the whole protein structure universe.

3.2.1 Objective Measures for Structure Alignment Quality

The most fundamental measures for structure alignment quality are derived from the spatial superimposition of the aligned structures' C_α traces. The commonly agreed on approach to compute this superimposition utilizes the MaxSub algorithm by Siew *et al.* [2000] to define the subset of aligned amino acid sites that maximizes the number of close C_α atoms in space. On this set the rotation is computed that complies with the Kabsch condition [Kabsch, 1976, 1978], minimizing the total cRMSD.
This two step procedure is needed because one is more interested in pointing out a common conserved core than getting the minimum total cRMSD value which can already take large values if only parts of the aligned structures are rather different, concealing the common core. The MaxSub step is therefore used to reject outliers that cannot be superimposed in a single rigid-body rotation. The routine is subject to a distance cut-off that defines when C_α atoms are close in space, a parameter usually set to $d_{\mathrm{th}}^{\mathrm{MaxSub}} = 4\text{Å}$ in accordance to $d_{\mathrm{th}}^{\mathrm{PSI}}$, and a seed length for the iteration used that is set to $L = 4$. The Kabsch transformation does not have any free parameters.
Once the superimposition is computed, it is easy to count all aligned C_α atoms closer in space than the cut-off value, which is chosen to $d_{\mathrm{th}}^{\mathrm{PSI}} = 4\text{Å}$ in order to ignore typical evolutionary variations in related structures, an example is shown in Fig. 3.5. Consecutive close sites build up the conserved core of the aligned proteins. Hence, the PSI quantifies the ratio of this conserved core in comparison to the shorter protein chain in the alignment. It is defined as

$$\mathrm{PSI} = \frac{\sum_{i=1}^{N_\mathrm{a}} \Theta\left(d_i - 4\text{Å}\right)}{\min\left(N_1, N_2\right)} \tag{3.8}$$

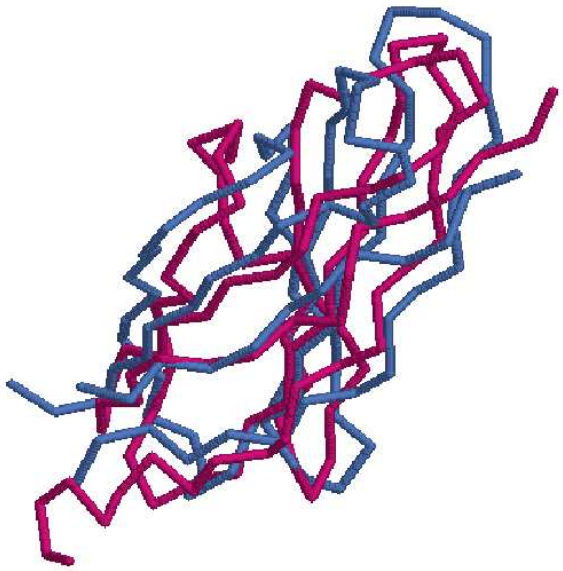


Figure 3.5: The spatial superimposition of the C_α traces for the alignment of d1cd9b2 vs. d1bpv__ as computed by SABERTOOTH is shown.

with N_a the number of aligned sites, $d_i = \left|\vec{x}^{(1)} - \vec{x}^{(2)}\right|$ the distance of aligned sites i, and N_1, N_2 the number of amino acids in the two chains.
An altered version of the PSI restricts the close sites considered to those that are member of a consecutive group of aligned sites that is longer than three, uninterrupted by gaps. We call this variant the relevant PSI or rPSI. It turned out that the PSI can differ strongly from the rPSI for alignments that have a large number of short aligned fragments, as shown by Teichert *et al.* [2007]. These fragments are most likely spurious hits that should not be considered when assessing alignment quality. Therefore, the rPSI is preferred for the analyses in this thesis.
The cRMSD

$$\mathrm{cRMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N_\mathrm{a}} \left|\vec{x}^{(1)} - \vec{x}^{(2)}\right|^2} \tag{3.9}$$

is especially meaningful when the full cRMSD can be compared to the $\mathrm{cRMSD}_\mathrm{core}$ that is restricted to the conserved core region of the protein. The cRMSD can take very large values even though the $\mathrm{cRMSD}_\mathrm{core}$ is negligible, which means that the aligned chains are indeed related but diverge outside the conserved core, given a large enough PSI value.
The contact overlap score q does not depend on a spatial superimposition and may therefore give additional information about the alignment quality outside the conserved core found in superimpositions. Its values furthermore depend on the details of the definition of the contact matrices used and have therefore to be considered

less fundamental than the PSI. This is especially true because the influence of the different parameters has not been investigated systematically yet.

$$q(C_{ij}, C'_{ij}) = \frac{\sum_{ij} C_{ij} C'_{ij}}{\min\left\{\sum_{ij} C_{ij},\, \sum_{ij} C'_{ij}\right\}} \tag{3.10}$$

Indices i and j sum over all components in the matrices reduced to aligned amino acid sites.
The sequence similarity (seqSim), even though derived from pure sequence information, can give additional information about alignment quality as stated before. If a high enough PSI value ensures that an alignment is structurally meaningful, seqSim can give evidence for details not seen in the PSI, since shifts by a single position may not change the PSI but the seqSim. Also here only relevant sites should be counted, in analogy to the rPSI, defining the rSeqSim. Sequence similarity values depend on the underlying substitution matrix, here BLOSUM62 was consistently used as it is the most universal compromise.

3.2.2 Similarity Significance Scores

All similarity measures are only applicable to pairwise comparisons since their numerical values depend on chain length, total number of contacts, or alike. A PSI of 80% could therefore mean to find eight close sites in an alignment of two structures with chain lengths 10 and 100, or to find 80 close sites in a comparison of two structures of length 100, leading to very different conclusions.
For many questions, like database clustering as mentioned before, all vs. all alignment of a set of structures is needed. A Z-score that can be used to assess the statistical significance of these alignments can be defined by comparing the PSI values found in an alignment to the expected statistical background distribution for unrelated alignments. This background will dependent on algorithm specific properties like mean number of aligned sites which can be quite different for different tools.
Assuming a normally distributed number of close sites in random alignments, the length dependence of the PSI can be stripped off by computing the mean PSI, $\langle \mathrm{PSI} \rangle$, and its standard deviation, σ_{PSI}, as functions of minimum chain lengths, yielding

$$Z_{\mathrm{struct}} = \frac{\mathrm{PSI} - \langle \mathrm{PSI} \rangle}{\sigma_{\mathrm{PSI}}}\,. \tag{3.11}$$

This Z-score can be used to compare alignments with each other. In the first example of above the PSI leads to an insignificant low Z-score value while it reveals

strong relatedness of the structures in the second example[2].
More due to historical reasons than solid reasoning the numerical values of the Z-score are assigned hierarchical levels of similarity, originally defined on the output of the DaliLite algorithm [Holm & Park, 2000, Holm & Sander, 1993]. These levels are related to the SCOP classification, assigning fold level similarity to aligned structures with $Z > 2$, superfamily level similarity for $Z > 4$, and family level similarity for $Z > 8$. For assigning two structures to the same SCOP class level only the secondary structure content of these structures needs to be similar.
Besides of the structural significance score derived from the PSI, an analogous procedure can be carried out for the sequence similarity measure. Although the seqSim in a structure alignment is much less meaningful than the PSI, it can provide additional information about the evolutionary distance of the proteins compared. These two measures could be vastly different due to the high degree of degeneracy in the sequence. Accordingly, the evolutionary significance score is defined as

$$Z_{\text{evol}} = \frac{\text{seqSim} - \langle \text{seqSim} \rangle}{\sigma_{\text{seqSim}}} \,. \tag{3.12}$$

3.2.3 Determining Z-scores for SABERTOOTH

The PSI values found when aligning unrelated structures trivially decrease for increasing chain lengths. The better the algorithm is able to distinguish between similar and dissimilar structures the steeper this decrease and the better a Z-score defined on PSI will perform.
Figure 3.6 shows the PSI values found for 31284 alignments from a set of unrelated structures plotted over chain length of the shorter chain. The set consists of the all vs. all combination of all structures from different superfamilies of the representative set of 29 superfamilies defined by Leo-Macias *et al.* [2005]. In addition, alignments with DaliLite or MAMMOTH Z-score larger than three were abstracted. The quantities $\langle \text{PSI} \rangle$ and σ_{PSI} needed to compute the Z-score can be fitted with power-laws. The resulting numerical values for the structural Z-score on PSI are

$$\langle \text{PSI} \rangle = a \cdot \min(N_1, N_2)^b = 554.975 \cdot \min(N_1, N_2)^{-0.721219} \tag{3.13}$$

$$\sigma_{\text{PSI}} = c \cdot \min(N_1, N_2)^d = 484.091 \cdot \min(N_1, N_2)^{-0.904381} \,. \tag{3.14}$$

The same power-law fit applied to sequence similarity, as shown in Fig. 3.7, results to the numerical values for the evolutionary Z-score on seqSim

$$\langle \text{seqSim} \rangle = -0.309173 \cdot \min(N_1, N_2)^{0.00051941} \tag{3.15}$$

$$\sigma_{\text{seqSim}} = 1.46487 \cdot \min(N_1, N_2)^{-0.416873} \,. \tag{3.16}$$

[2]for the parameters found for SABERTOOTH the Z-scores are -0.43 and 8.3, respectively

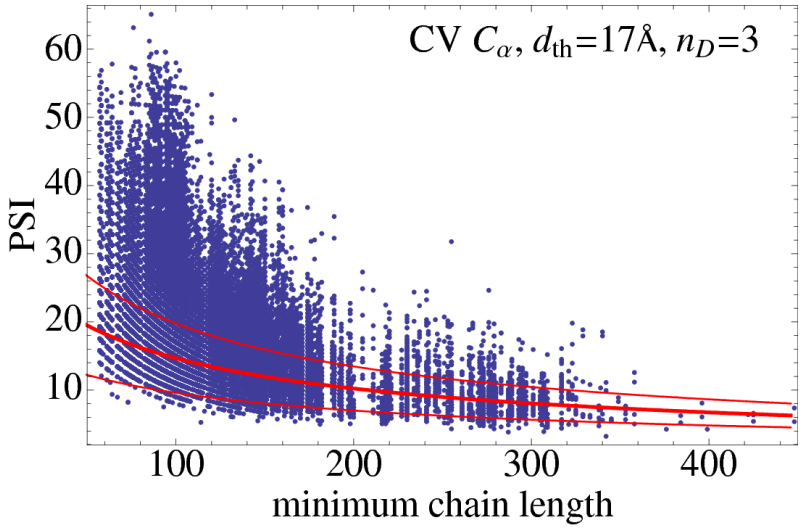


Figure 3.6: The scatter plot shows the PSI over shorter chain lengths for the 31284 alignments of unrelated pairs in the Z-score set. The thick line marks the $\langle \text{PSI} \rangle$ while the thin lines mark $\langle \text{PSI} \rangle \pm \sigma_{\text{PSI}}$.

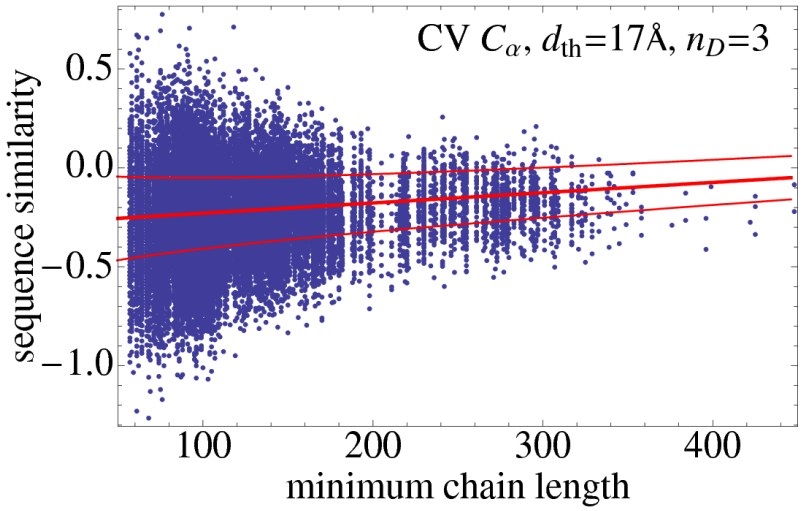


Figure 3.7: The scatter plot shows the seqSim over shorter chains length for the 31284 alignments of unrelated pairs in the Z-score set. The thick line marks the $\langle \text{seqSim} \rangle$ while the thin lines mark $\langle \text{seqSim} \rangle \pm \sigma_{\text{seqSim}}$.

3.3 Comparison to References

Three different kinds of tests are important when assessing the quality of an alignment algorithm. The main obligation is, of course, that structural similarities are recognised as accurately as possible. Furthermore, structures with relevant similarity should be clearly distinguishable from dissimilar structures using a statistical significance score like the Z-score defined above. If an algorithm is too tolerant when aligning rather different structures this is reflected in an imprecise significance score, unusable for classification purposes. Although these demands are certainly entwined, they are not synonymous and separate tests are required to assess these skills. In addition, short computation times are crucial in order to permit large scale application of the tool in which potentially millions of alignment runs are called for. The results of both accuracy and classification performance tests strongly depend on the extent of similarity in the alignments of the test set under investigation. Therefore, it is advised to carry out separate tests for different levels of similarity. The database for the Structural Classification of Proteins (SCOP) by Murzin *et al.* [1995] can be consulted to construct such sets. In SCOP structures are assigned to similarity clusters that constitute a hierarchical classification scheme. The four major similarity levels in SCOP are class, fold, superfamily, and family going top-down to more similar structures. While the class level is simply defined by secondary structure content without describing any evolutionary relationship, the other three levels are traceable to alignments.
At the most similar level, structures of the same family have a clear evolutionary relationship, they most likely share the same function and relationships are already evident from the sequence for most of the cases. The superfamily level in the SCOP hierarchy comprises structures for which low sequence similarities disqualify sequence based analyses but structural or functional features suggest a common evolutionary origin. Comparison of superfamily related pairs is demanding also for structure alignment tools. Definitive answers about specific relations can only be derived through a combination of significance score and additional information about function or evolution, limiting automated analyses. Fold level related structures share similar motifs of secondary structure but large portions of the structures may differ in secondary structure and overall conformation, evolutionary relation is unspecified after all.
Computation speed can be different for similar and for dissimilar test sets as well. Algorithms that rely on iterating a heuristic optimization scheme may converge much faster on related than on unrelated pairs. In contrast to that, tests run prior to the actual alignment routine can be used to save time by sorting out very different pairs beforehand. Consequently, speed tests should be carried out on two different test sets, one with fairly similar pairs and one with unrelated pairs.

3.3.1 Similarity Recognition at different evolutionary Distances

Structure alignment tools are designed to detect similarities in pairs of structures. The specification of algorithm and data used will be reflected in the performance achieved. Three test sets of alignments were defined with alignments from the fold, superfamily, and family levels of the SCOP classification (version 1.73) to find out how accurately structural similarities are detected. A percentage of the examples from the all vs. all combination of all structures of the respective levels of similarity was selected randomly to compile these test sets, discarding pairs that also fall into the same cluster of the underlying level. Also chains with less than 30 amino acids were rejected. When, for example, assembling the superfamily test set, 4.5% of all possible pairs from the same superfamilies that are from different families form the test set.
For all test sets the best alignments of six freely available established reference tools were selected by picking out the candidate with the largest rPSI and rSeqSim values for each alignment, in order to get the best possible reference sets, called best-of sets in the following.
Most prominent members of the reference tools are DaliLite and MAMMOTH. DaliLite is very widely used for more than 16 years now and a quasi-standard in the field. It is based on the direct alignment of distance matrices, yielding very good results. MAMMOTH is used by the CASP assessors [Tramontano, 2007] and in the folding@home project [Larson *et al.*, 2002, Shirts & Pande, 2006]. Other references are CE [Shindyalov & Bourne, 1998], TM-align [Zhang & Skolnick, 2005], SHEBA [Jung & Lee, 2000], and MAMMOTH-multiple [Lupyan *et al.*, 2005]. CE starts from local structure fragments that are extended to form the final alignment. TM-align is based on optimizing the so-called TM-score [Zhang & Skolnick, 2007], a significance score derived from the cRMSD. MAMMOTH-multiple also computes high quality pairwise alignments that are partly different from those output by the actual pairwise version. Both algorithms are based on profiles build up from angles between consecutive C_α atoms.
All measures on display (except of the significance scores) were re-computed from the alignment strings output by the programs. This was done to compile the best-of sets and also for all measures shown for comparison reasons. It was advised to do so to make sure that all numerical values are comparable and the results are not influenced by simple deviations in e.g. specification of the superimposition, cut-offs and so on. The routine to carry out this task consists of the following steps:

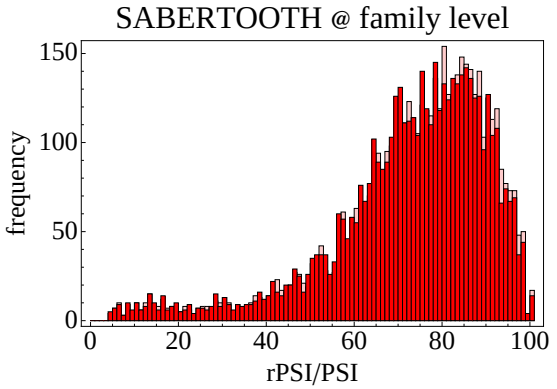
- read the external sequence alignment string as computed by the reference tool
- read the respective coordinate files
- align the sequences given from the external tools with the sequences extracted from the coordinate files to recover the correct sequence to coordinate mapping

SCOP level	Coverage	# of alignments
family	611/1611	5014 (≙11.0%)
superfam.	244/1008	4981 (≙4.5%)
fold	83/654	4737 (≙2.3%)

Table 3.2: The table lists the coverage of three major levels of the ASTRAL40 selection of the SCOP database that consist of more than one member. The number of alignments in the respective sets, and the percentage of all possible alignments on this level are stated. It shows that large part of all clusters on each level are represented, the seven clusters of the class level are fully covered.

- perform a Kabsch transformation based on the MaxSub set of aligned pairs
- compute cRMSD and PSI (rPSI) measures on the spatial superimposition
- compute seqSim (rSeqSim) from the sequence alignment
- compute contact overlap

Not all reference tools output results for all alignment examples which is partly due to the fact that some tools suppress to output the actual alignment for examples they assigned too little significance, partly due to simple bugs. Even if only one example is missing, the respective alignment is abstracted from the set. This reduces the size of the fold set by 7.5% and the superfamily and family sets by 5%, mainly due to the large number of missing results from DaliLite. The final test sets cover a large fraction of the clusters of the ASTRAL40 selection of the SCOP database that consist of more than one member, as shown in Table 3.2. ASTRAL40 stores PDB-style coordinate files of all domains defined in SCOP that have pairwise sequence identities below 40%.
SABERTOOTH is not expected to perform as well or even better than the best-of reference but at least on the SCOP family level it is very close to this optimum. The main difference here is shown in the low rPSI tail in Fig. 3.8. SABERTOOTH could not assign rPSI values greater than 40% for 310 alignments compared to only 76 cases for the reference. Besides of that the difference histogram in the same figure proves that for most of the alignments SABERTOOTH and the best-of reference are assigned very similar rPSI values. Also mean cRMSD values confirm this picture. Looking at rSeqSim SABERTOOTH performs significantly better than even the best-of reference which might be due to the sequence dependent term that is included in SABERTOOTH's cost function, while all references except of SHEBA do not use the sequence as an additional source of information, even though this can



	SABERT.	best-of set
rPSI	70.2% (69.3%)	74.2%
rSeqSim	0.355 (0.300)	0.287
cRMSD	5.10Å (5.92Å)	4.46Å
rPSI < 40%	310 (383)	76

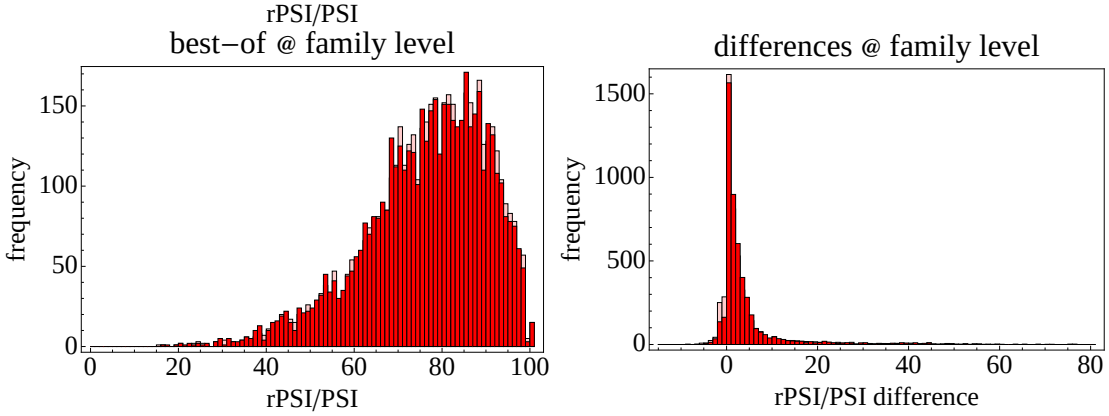


Figure 3.8: The rPSI distributions found over the family level test set are shown for SABERTOOTH and the best-of reference. The difference histogram proves that SABERTOOTH highly agrees with the reference in the similarities found. Mean rPSI and cRMSD values are comparable while SABERTOOTH has an error rate of less than 5% in which it cannot find a good alignment. rSeqSim in contrast is higher than even in the best-of reference. Data given in brackets refer to EC based alignments.

be done without disturbing the actual structure alignment, as demonstrated by the results.

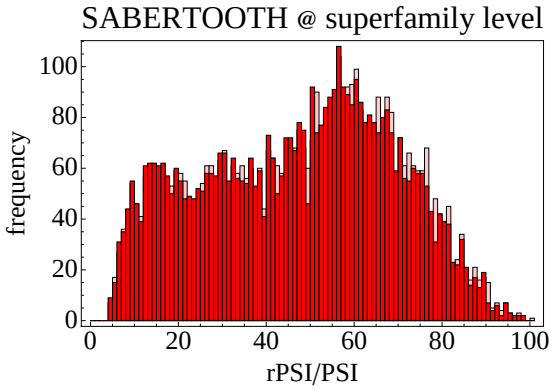
The first astonishing finding when looking at superfamily and fold level results is that these test sets are of the same structural diversity in the margin of errors. This is in contrast to what one would expect from a structural classification database. In fact, superfamily and fold levels can be distinguished by the significantly higher sequence similarity found in the superfamily related alignment suggesting that the definition of the levels mainly relies on evolutionary relatedness, probably evaluated by sequence alignment tools.

The comparison of SABERTOOTH's performance to the best-of reference, as shown in Fig. 3.9 reveals a slight degradation in alignment quality when going to more different structures. On both superfamily and fold level there is a gap of around ten

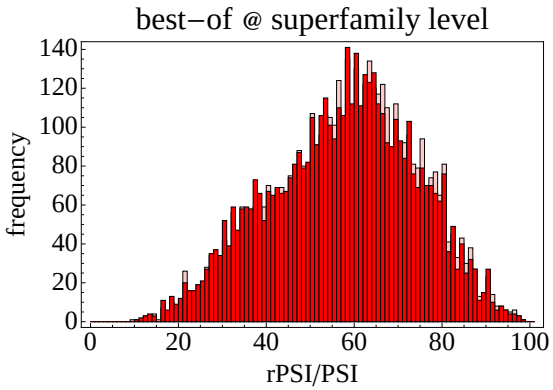
percentage points in rPSI along with larger mean cRMSD values. The difference histograms, still peaked at very small values, clearly show a growing tail of assumedly wrong alignments for growing structural/evolutionary distance. Despite of that, the sequence similarity found is consistently larger for SABERTOOTH's alignments than for the reference sets again underpinning the use of sequence information for the alignment. This result is even more encouraging, taking into account that optimizing for structural and sequence similarity at the same time are partly conflicting obligations, which explains part of the gap in structure alignment accuracy.
From a structural point of view the alignment quality of SABERTOOTH is very close to the best-of reference if the structural distance is not too large, whereas SABERTOOTH is much more precise from a sequence based point of view over the whole spectrum of diversity.
The results for the EC based alignments by SABERTOOTH are very similar in quality. The EC performs consistently slightly worse and its quality degrades slightly faster with structural distance. Nevertheless, considering the very different definition of the profile used, the overall picture agrees very well.

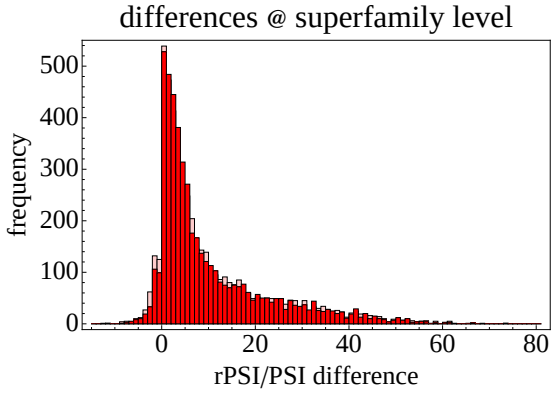
3.3.2 Comparison with established Structure Alignment Tools

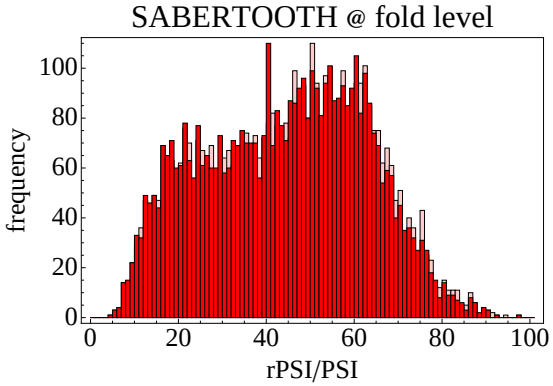
In addition to the comparison with the best-of set, also the performance of the tools themselves is of interest. On the one hand, possible excellences and weaknesses are revealed and, on the other hand, a detailed comparison of the tools amongst each other and with the one presented here gives insight in the quality of currently available tools.
To do so, the measures introduced before are computed for all tools and over all test sets. The resulting mean values are summarized in Table 3.3 on page 48. The test sets are identical to those used in the preceding section, allowing for direct comparison with SABERTOOTH.
The widening gap in alignment quality in comparison to the best-of reference that was already seen in the comparison with SABERTOOTH is found for all reference tools in approximately the same severity. DaliLite sticks out with its ability to assign proper alignments in almost all cases, it furthermore reaches the highest mean rPSI and contact overlap values making it the most accurate tool assessed here. Results shown for DaliLite may slightly overestimate its performance since the test sets are biased in its benefit. Many examples where sorted out in the construction of the sets because DaliLite did not output results. For a large fraction this could have happened due to the fact that DaliLite could not find proper similarities, masking its true error rate. TM-align is second best in this discipline and stands out when looking at the low mean cRMSD values reached.



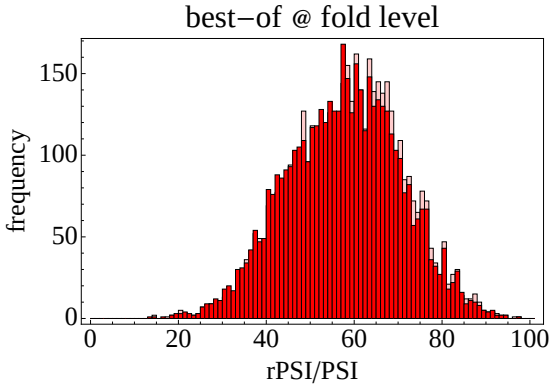
	SABERT.	best-of set
rPSI	43.6% (41.4%)	53.3%
rSeqSim	-0.138 (-0.250)	-0.261
cRMSD	10.37Å (12.56Å)	7.57Å
rPSI < 30%	1291 (1481)	289







	SABERT.	best-of set
rPSI	41.2% (40.1%)	53.6%
rSeqSim	-0.309 (-0.408)	-0.462
cRMSD	9.43Å (10.99Å)	6.66Å
rPSI < 20%	577 (572)	8



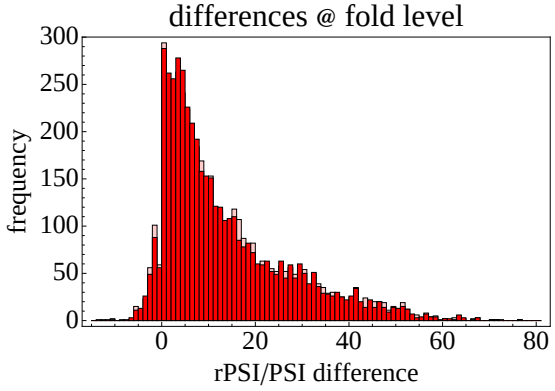


Figure 3.9: The rPSI distributions found over the superfamily and the fold level test sets are shown for SABERTOOTH and the best-of reference. The difference histograms show a growing tail of erroneous alignments by SABERTOOTH while mean rPSI and cRMSD values are still only about 10% points below the reference. Although negative for these sets, better rSeqSim are found for the SABERTOOTH alignments than for the best-of reference. Data given in brackets refer to EC based alignments.

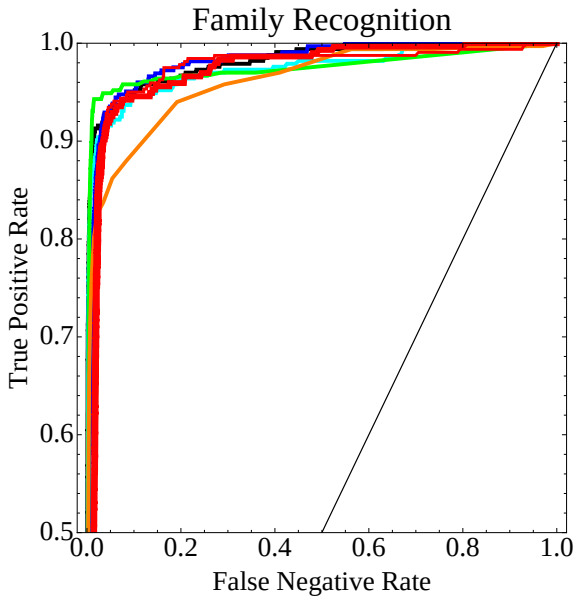
SHEBA clearly performs worst over all test sets, its alignment quality degrades faster for growing structural distance. CE, in contrast, achieves better results in comparison to the other competitors on the fold level than on the family level.
SABERTOOTH performs comparably well to the references but the sequence similarities found in all the test sets is significantly higher. For farther structural distances SABERTOOTH's error rate grows a bit stronger than for the best tools of the reference set.

3.3.3 Structural Classification Abilities

Not only an algorithms ability to discover similarities is of importance but also its capability to clearly distinguish similar from dissimilar pairs of structures. A structural significance score, as the one defined in Section 3.2.2, can be applied to assess this capability by testing the agreement of this score with a structural classification database. A suitable test set consists of a mixture of related and unrelated alignments. Here 498 domains from 97 different folds were randomly selected from ASTRAL40 (version 1.73) [Chandonia *et al.*, 2004]. The different folds are represented in the set relative to their number of members in ASTRAL40.
All tools were then used to compute the 123753 all vs. all alignments of the structures in the test set and output their respective significance scores. This so-called similarity matrix containing all pairwise scores is then input to display the data in the form of ROC-plots[3] in order to graphically quantify the agreement of the significance scores with the SCOP classification. In a ROC-plot the cut-off for the respective score is shifted from its minimum to its maximum value, counting the number of alignments with $Z > Z_{\text{cut-off}}$, the positives P, and $Z \leq Z_{\text{cut-off}}$, the negatives N in each step. If a positive P is also in the same fold, superfamily, family of the classification, respectively, the assignments agree and the example is counted as a true positive TP, if not the alignment tool and the classification disagree and the example is counted as a false positive FP. The same is done for the negatives: true negatives TN agree with the classification, false negatives FN are those that overlook similarity assigned by the classification.
To draw the ROC-plot curve the ratios of TP/P are plotted over FN/N. The further the curve falls to the left in the plot the less negatives are assigned high values of similarity, while pushing the curve up means that more true positives are recognized correctly. This can be understood as measuring coverage and sensitivity. The larger the area under the curve (AUC), the better the agreement between significance score and classification, while the diagonal would result from random decisions.

[3] *ROC* stands for 'Receiver Operating Characteristic' and is borrowed from the field of signal detection theory

The score itself is only used in parametrizing the plot but is not explicitly drawn. This is important because the scores of different programs might be defined with very different absolute values that may not be comparable, otherwise. This is also one reason why no best-of compilation can be introduced here, as done for the alignment quality assessment. The other reason is that the scores may be based on very different measures and use involved fitting procedures that cannot be implemented without in-depth knowledge of the program.
Due to inherent ambiguities in a classification and maybe simple mistakes the classification database is not perfect and deviations from AUC $\equiv 1$ are expected already on family level. For superfamily and fold levels pure structural information might not be sufficient to decide for every relationship and further expert knowledge about evolutionary or functional relation is needed, limiting the expected quality. The same kind of analysis on the SCOP class level is not meaningful since it could be carried out simply by analysing the secondary structure content, rather than a proper alignment.
Figure 3.10 shows separate ROC-curves for the three SCOP levels fold, superfamily, and family for all tools tested in this work except of SHEBA that does not output a significance score, together with a table listing the respective AUC values. On the family level all tools, maybe except of CE, work similarly well. This is not surprising because mean pairwise similarities of rPSI $\sim 70\%$ can be expected for this level, as seen in Section 3.3.2. The superfamily and fold levels are more demanding and DaliLite achieves much higher accuracy and coverage here than all other tools. This could partly be expected from DaliLite's high quality alignments that only slightly degrade in quality when going to less related structures. Nevertheless, the extent of advantage in this test is noteworthy. On superfamily and fold levels CE and also MAMMOTH-multiple perform clearly worse than the rest. TM-align's ROC-curve reveals a slightly different characteristic that might be caused by the TM-score that is used to measure significance in this case. The worse coverage is compensated by better accuracy, which reflects that TM-align shows the lowest cRMSD values assigned in quality assessment.
The performance of SABERTOOTH is in the upper midfield. It is interesting to note that EC based alignments lead to better classification abilities on the family level, even though alignment quality is slightly worse there. The performance decreases faster through superfamily and fold level than for the CV based program, which could be expected from the same tendency found for alignment quality.



AUC in %	fa	sf	fl
SABERT. CV	96.91	91.95	91.39
SABERT. EC	97.21	91.61	90.89
CE	96.01	88.56	85.48
DaliLite	97.40	96.24	95.20
MAMMOTH	98.08	91.67	92.08
MAMM.-mult	97.32	89.15	88.00
TM-align	98.13	91.27	91.40

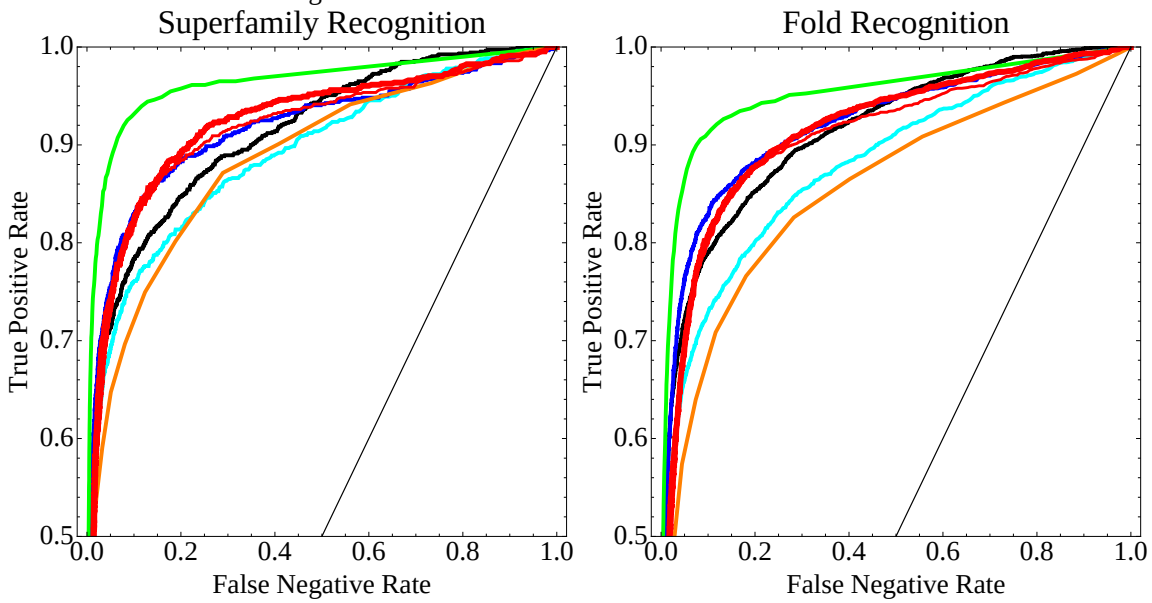


Figure 3.10: ROC-plots for the SCOP levels family (fa), superfamily (sf), and fold (fl) are shown. DaliLite performs much better in the classification assignments than all other tools. CE and MAMMOTH-multiple perform worse than the other tools that are of roughly comparable quality. SABERTOOTH's performance is very similar for CV and EC based alignments (thick and thin red curves) but the quality of the EC decreases faster for more distant alignment, as expected from the alignment quality assessment. Note that the ordinates start at value 0.5.

3.3.4 Computation Speed Comparison

Computation speed is a trivial but, nevertheless, important attribute of an alignment algorithm. The best tool cannot be used in practical contexts if it is just too slow to be applied to the alignment set in question. When e.g. computing the similarity matrix for database clustering, all vs. all alignment of all structures in this particular database is asked for. The PDB database currently holds about 125000 protein chains so that about $8 \cdot 10^9$ alignments were necessary. It is clear that even with a large number of processors and very efficient tools this is nearly impossible.
Key to fast algorithms is a reduced structure representation. It is an advantage of SABERTOOTH to use vectorial profiles in comparison to, e.g. DaliLite that relies on distance matrices making it by far the slowest tool assessed here. But also the vectorial profiles themselves can be complicated and very costly to compute. For instance, for SHEBA's environmental profiles detailed analysis of secondary structure is needed, also SABERTOOTH's alternative version based on the EC profile needs diagonalization of the contact matrix, a fairly slow algorithm of complexity $\mathcal{O}(n^3)$ with matrix dimension n.
SABERTOOTH itself is faster on related than on unrelated pairs. For the ideal case of identical structures, the alignment matrix would just be crossed on the diagonal but already for slightly dissimilar structures the profile alignment step covers the whole matrix. Here, the post-processing step is accelerated by assigning a smaller number of allowed components in alignment matrix A_{ij}, reducing the number of steps in the second run of Dijkstra's shortest path algorithm.
In Fig. 3.11 scatter plots of the algorithms runtimes[4] are plotted over problem size, i.e. the product of chain lengths in the alignment, for the SCOP family test set, that was also used for alignment quality assessment, and a set of 4465 alignments constructed by all vs. all combination of 95 randomly selected structures from different folds in the SCOP classification. The total runtimes over the test sets are shown in units of the total runtime of MAMMOTH, the fastest tool assessed. The MAMMOTH algorithm shows a strictly linear dependence of runtime on problem size independent from the structural distance in the alignment making it a stable reference.
Measuring runtimes on two different test sets is motivated by two typical tasks: Firstly, when an algorithm is used for small numbers of alignments in a web server or directly at the console, response times should be short. For this task the pairs are usually manually selected and some previous knowledge exists. Especially when computing an evolutionary common core of a set of structures known to belong to the same family, many alignments of related pairs are asked for. Secondly, when

[4]Timings are taken on an Intel(R) Xeon(TM) CPU with 2.80GHz.

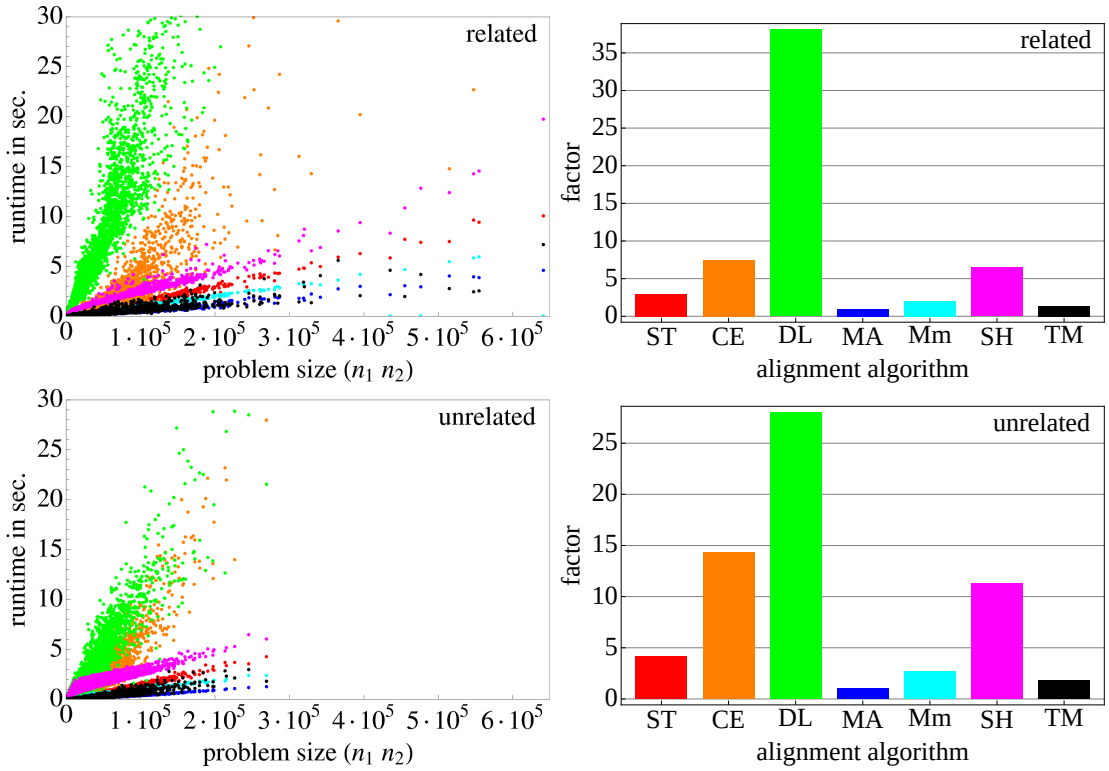


Figure 3.11: The scatter plots on the left hand side show runtimes over product of chain lengths. On the right hand side total runtimes are plotted in units of MAMMOTH's runtime. Upper plots refer to the family level test set used already before for quality assessment, containing related structures. Lower plots refer to unrelated structures. The ordinate in the scatter plots is chopped, DaliLite needs 328 sec. to execute the largest example in the related set.

huge databases are scanned for related structures, the abundant majority of alignments that have to be performed in an all vs. all set are between unrelated pairs. The slowest tool by far is DaliLite. Through some heuristics that sort out unrelated pairs even before the whole alignment run is carried out, it is faster on unrelated pairs than on related pairs. All other tools are faster on the related set than on the unrelated set, even though the difference is minor for MAMMOTH-multiple and TM-align. SABERTOOTH's runtime increases in the mean by roughly one third when doing unrelated alignments, CE's and SHEBA's runtimes nearly double. Only the runtimes of DaliLite and CE seem to increase faster with problem size than linear.

SCOP family level test set, 5014 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 40%
best-of ref.	74.2%	4.46Å	0.612	0.287	76
SABERTOOTH	70.2%	5.10Å	0.584	0.355	310
CE	67.2%	4.21Å	0.546	0.264	500
DaliLite	73.0%	5.27Å	0.621	0.286	113
MAMMOTH	71.1%	5.47Å	0.603	0.269	256
MAMMOTHmult	70.8%	6.49Å	0.596	0.233	307
SHEBA	64.7%	5.44Å	0.549	0.309	408
TM-align	71.0%	2.96Å	0.595	0.317	164

SCOP superfamily level test set, 4981 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 30%
best-of ref.	53.3%	7.57Å	0.417	-0.261	289
SABERTOOTH	43.6%	10.37Å	0.358	-0.138	1291
CE	42.9%	7.61Å	0.327	-0.243	1225
DaliLite	52.0%	9.15Å	0.428	-0.281	369
MAMMOTH	45.1%	11.52Å	0.390	-0.327	1134
MAMMOTHmult	43.4%	14.16Å	0.382	-0.422	1338
SHEBA	37.2%	9.91Å	0.305	-0.240	1580
TM-align	48.0%	4.14Å	0.378	-0.198	662

SCOP fold level test set, 4737 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 20%
best-of ref.	53.6%	6.66Å	0.413	-0.462	8
SABERTOOTH	41.2%	9.43Å	0.336	-0.309	577
CE	43.1%	6.38Å	0.320	-0.437	257
DaliLite	51.9%	8.22Å	0.427	-0.481	20
MAMMOTH	44.0%	10.24Å	0.367	-0.503	260
MAMMOTHmult	41.9%	12.73Å	0.351	-0.592	453
SHEBA	35.4%	9.43Å	0.279	-0.373	732
TM-align	47.5%	4.11Å	0.369	-0.385	89

Table 3.3: The table shows the results for the assessment of structure alignment accuracy on the three test sets for different structural distances. rPSI refers to the relevant PSI, cRMSD measures the mean square distance of aligned sites, contact overlap is computed on heavy-atoms contact matrices with $d_{\text{th}} = 4$Å and $n_{\text{D}} = 3$. rSeqSim measures the relevant sequence similarity based on a BLOSUM62 matrix. In the last column the number of alignments below the respective rPSI cut-off is counted.

4 Vectorial Description of Protein Sequence

Protein structure alignment is possible with good results using connectivity based profiles, as discussed in the preceding chapter. These profiles are of vectorial form just like the protein's sequence and analogies are reaching farther: Already the very simple hydrophobicity profile (HP), a 20 parameter vectorial description of the sequence with only one fixed value per amino acid type, has significant correlation with the structural profiles PE and EC. Both these profiles can be used to derive site-specific amino acid distributions, predicting the compatible sequences for a given structure, as shown by Bastolla *et al.* [2008, 2006].
Driven by this intriguing relation of sequence and structure, a prediction of the structural profile using only sequence data as input, a highly active field of research today, seems promising. Subsequently, these predicted structural profiles could be plugged in into the alignment framework originally developed for structure alignments in order to perform sequence alignments. It was demonstrated in the preceding chapter that profiles as different as the CV of the C_α trace and the EC of heavy-atoms can be used in the same alignment scheme with good results, raising the hope that also a predicted profile, even if it is not perfectly well predicted, can give good results.
At first in this chapter, the relation of protein sequence and structure is discussed. Then a standard artificial neural network approach is adapted to predict the structural profile from the sequence that was used before to perform structure alignments, namely the contact vector CV with distance cut-off $d_{\text{th}} = 17$Å and $n_{\text{D}} = 3$. This work was carried out as part of the diploma thesis of Jonas Minning. Thereafter, the predicted profiles are compared to those computed directly from protein structure to evaluate the quality of the prediction.
In the next chapter these predicted profiles are used to perform sequence alignments that are then assessed using the same systematic applied to assess the structure alignments in the preceding chapter.

4.1 The Sequence/Structure Relation

That protein sequence and structure are deeply related is obvious envisioning the fact that a strand of DNA encodes a specific protein sequence which, in turn, is expressed to a specific protein structure. This mechanism is fundamental for every organism as a protein's function is based on its precise structural properties. In contrast to that, the other way around is not that strictly defined and many different sequences comply with a given structure. In the course of evolution, every site in a protein sequence is being substituted over and over again with similar amino acids, mostly without substantially changing the structure. In fact, two sequences with unrecognizably low sequence similarity can encode for nearly identical structures.
This is the underlying reason to look for sequence representations that employ physical or chemical properties of the amino acids that are more stable under amino acid substitution. Many of these definitions exist ranging from hydropathy values and surface accessibility to secondary structure propensities, to name only few.
A prominent representative of sequence representations is the hydrophobicity profile (HP). It is one example of the class of 20 parameter approximations, listing one typical amino acid hydrophobicity value in this case. To get the HP the hydrophobicity values of a given sequence are listed according to the residue types. It gets intuitively clear that this hydrophobicity profile should be correlated with the structural effective connectivity EC for folded protein structures in aqueous solution, when considering that more hydrophobic amino acids are folded inside the protein structure to build up the hydrophobic core, while at the protein's surface hydrophilic amino acids dominate.[1] The hydrophobic core is also the place where contact density and, hence, effective connectivity is higher than elsewhere in the structure.
Yet for this very basic model that ignores any correlation within the sequence, the chain length weighted Pearson correlation coefficient over a representative subset of the PDB with 10892 chains (PDB clusters 2008-08-15, 50% sequence identity, rank 1) yields $r(\mathrm{HP}, \mathrm{EC}) = 0.432$.
Bastolla *et al.* [2006] show that this sequence to structure relation can be exploited to determine site-specific amino acid propensities for a given structure starting from the PE representation of structure with good agreement with experimental data. Furthermore, the optimal hydrophobicity profile $\mathrm{HP}_{\mathrm{opt}}$ is introduced that can be regarded as the central direction in sequence space around which the evolution of the sequence librates for a given structure. In this picture it gets clear that the vector spaces for the description of sequence and structure are unified, a fact that ex post

[1] In fact, different treatment is needed for membrane proteins that are embedded in aliphatic environments.

allows to compute e.g. correlation coefficients of sequence and structure representations.
In Bastolla *et al.* [2008] it is shown that the normalized $\mathrm{HP}_{\mathrm{opt}}/\langle \mathrm{HP}_{\mathrm{opt}} \rangle$ is a member of the GEC family of profiles for the choice of $B_{\mathrm{HP}} = \langle \mathrm{HP}^2_{\mathrm{opt}} \rangle / \langle \mathrm{HP}_{\mathrm{opt}} \rangle^2$ setting the former reasoning on formal grounds.

4.2 Predicting structural Profiles using a Neural Network Approach

Many groups and worldwide projects like CASP[2] [Tramontano, 2007] are working in the field of protein structure prediction from the sequence, as it is one of the supreme disciplines of protein science, largely unsolved until today. The task of predicting an approximate structural profile, in contrast, is much less challenging. Nevertheless, already the very unpretentious HP is applicable for selected analyses. More elaborated techniques that exploit correlations between the amino acids of the protein sequence are already employed with good results to predict a number of structural properties, as domains, secondary structure propensities, solvent accessibility, and even residue contact matrices [Jones, 1999, Kinjo & Nishikawa, 2005, 2006, Vullo *et al.*, 2006].
A general idea when using sequence data is to search a database to identify as many homologous sequences as possible in order to recover information about the whole family a given query sequence is member of. Collecting statistics over a huge multiple alignment of sequences, a site-specific probability of each amino acid type can be derived experimentally giving insight into the evolutionary stability of each site in the sequence. A widely used tool to align a query sequence to a database is the PSI-BLAST algorithm[3] by Altschul *et al.* [1997]. It performs pairwise BLAST alignments of a query sequence to all sequences in a database using a substitution matrix to measure similarity. In each iteration the database is filtered for compatible sequences and the initially used all-purpose substitution matrix is refined to better fit the specific query sequence. A typical sequence database used for this task consist of around 5–10 million sequences, enough to give good statistics if the family of the query sequence is well represented in the database. Alongside with the final alignment PSI-BLAST also outputs the refined substitution matrix, the so-called Position Specific Scoring Matrix (PSSM). This PSSM tabulates the probability to find each amino acid type in each sequence site as a 20 times chain length sized matrix.

[2]CASP is an acronym for *Critical Assessment of Techniques for Protein Structure Prediction.*
[3]The name *PSI*-BLAST derives from Position Specific Iterative, not to be confused with Percentage of Structural Identity.

In Jones [1999] an artificial neural network is trained with PSSMs to predict secondary structure propensities from the sequence with very good results. The same idea can be employed to predict all kinds of structural profiles including contact vectors, similar to what is done by Kinjo & Nishikawa [2005].
Since high quality structure alignment was possible with very different structural profiles, it is expected that sequence alignments will be feasible, even though the profile prediction is not perfect and the profiles themselves will be far from equivalent to the actual structure by construction.

4.2.1 Implementation to predict structural Profiles

The implementation of the neural network to predict structural profiles is comprised of three layers: Input, hidden, and output layer. The input layer has $21 \cdot 15 = 315$ neurons, $20 + 1$ accounts for the 20 different amino acid types with an additional terminus marker. The window over which sequence correlation is expected is 15 amino acid sites wide. The hidden layer has 40 neurons that feed the output neuron. Window size and number of hidden layer neurons are free parameters, in principle, the particular choice made here is based on extensive brute-force testing.
In the training phase online-learning and early-stopping is used in order to improve convergence in minimizing the RMSD of the prediction in respect to the exact profile. The training of the network was conducted over a set of 1500 globular chains from a non-redundant subset of the PDB with chain lengths between 30 and 300 amino acids that was divided into a training and a validation set. For each sequence in the training set the exact structure derived profile and the PSSM computed by PSI-BLAST were given as input.
The network algorithm is mostly independent of the specifications of the predicted profiles but we aim here to predict the profile used for structure alignments before, i.e. the contact vector on the C_α trace with $d_{\text{th}} = 17$Å and three excluded diagonals. Since the training scheme minimizes RMSD, mean value and standard deviation of the resulting prediction are not determined. While the mean can be normalized to one simply through dividing by the profile's mean as done for the structural profiles, the variance is virtually unknown and therefore especially predicted using a scheme based on an ansatz made by Kinjo & Nishikawa [2005]. Sequence information enters through computing the 20 mean values, one for each amino acid type, over the given sequence-specific PSSM. The length dependence of the variance is fitted by introducing a length dependent term that describes the scattering of the variance and by modelling the length dependent mean value of the variance as a power law. The functional form that is being used to predict the contact vector's variance $\tilde{\sigma}$

reads

$$\tilde{\sigma} = \left(\sum_{i=1}^{20} \langle \mathrm{PSSM}_i \rangle \cdot f_i \right) \cdot l^a + b \cdot l^c \tag{4.1}$$

with the amino acid specific fit parameters f_i, and parameters a, b, and c for the length dependence. Furthermore, a lower bound value $\tilde{\sigma}_{\min} = 0.05$ is used to suppress artificially low variance values.

4.2.2 Prediction Quality

Prediction quality is evaluated over a set of protein structures that is independent from the training set. The results are compared to those found for the meanCV profile that just lists mean CV components per amino acid. The meanCV is introduced here as the most simple approximation of the CV that can be derived from the sequence, as it consists only of the mean component values per amino acid type found in a database search. Thus, it is a 20 parameter prediction of the CV and should therefore be roughly of the quality as the HP compared to the EC since both do not account for any inner sequence correlation.
In fact, the HP is expected to correlate better with the EC than the CV with the meanCV due to two reasons: First, sequence to structure correlation is generally higher for profiles derived from heavy-atoms matrices than for backbone based ones, and, second, already the meanEC, analogue in definition to the meanCV, is slightly less correlated with the EC than the HP.
The 20 mean values for the meanCV were computed over a non-redundant subset of the PDB with 50% sequence identity cut-off, similar to the one used to train the neural network. The numerical values used can be found in Table A.1, alongside with the hydrophobicity values used for the HP.
To assess the prediction quality of the predCV, Pearson's correlation coefficient is computed as

$$r(\mathrm{predCV}, \mathrm{CV}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathrm{predCV}_i - \langle \mathrm{predCV} \rangle}{\sigma_{\mathrm{predCV}}} \cdot \frac{\mathrm{CV}_i - \langle \mathrm{CV} \rangle}{\sigma_{\mathrm{CV}}} \tag{4.2}$$

over all 9420 chains of ASTRAL40 (version 1.73), with $\langle . \rangle$ the mean and $\sigma_.$ the standard deviation. It turns out that the length weighted mean correlation $r(\mathrm{CV}, \mathrm{meanCV}) = 0.290$ is rather weak, while the prediction correlates as strong as $r(\mathrm{CV}, \mathrm{predCV}) = 0.717$, as shown in Fig. 4.1. The variance fit using the ansatz Eq. (4.1) leads to a rather weak gain in prediction accuracy, as depicted in Fig. 4.2 where the variance as output by the network is plotted over the target variance in comparison to the variance after fitting. The correlation coefficient of the variances over the 9420 structures of ASTRAL40 improves from $r(\sigma_{\mathrm{CV}_{\mathrm{target}}}, \sigma_{\mathrm{CV}_{\mathrm{ANN}}}) = 0.235$ to $r(\sigma_{\mathrm{CV}_{\mathrm{target}}}, \sigma_{\mathrm{CV}_{\mathrm{fit}}}) = 0.385$.

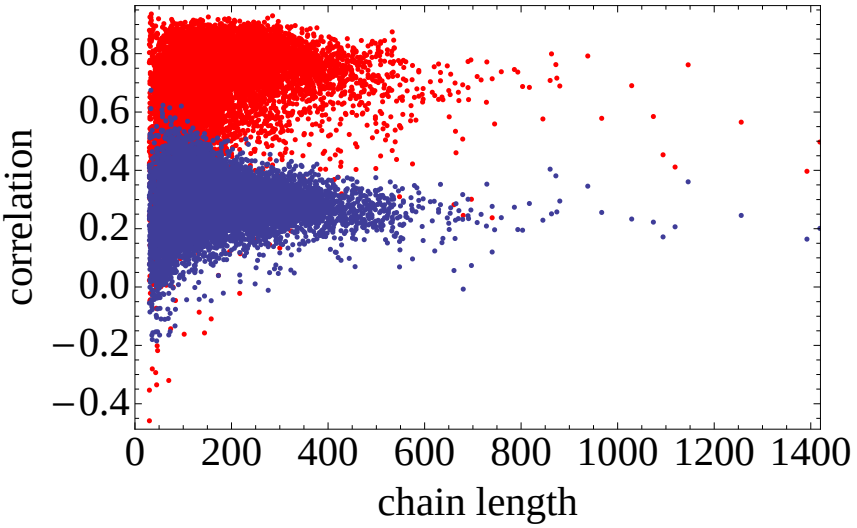


Figure 4.1: The scatter plot of the correlation of the predicted contact vector with the target is shown (red). The length weighted mean correlation coefficients over all 9420 chains of ASTRAL40 (version 1.73) are $r(\mathrm{CV}, \mathrm{predCV}) = 0.717$ and $r(\mathrm{CV}, \mathrm{meanCV}) = 0.290$.

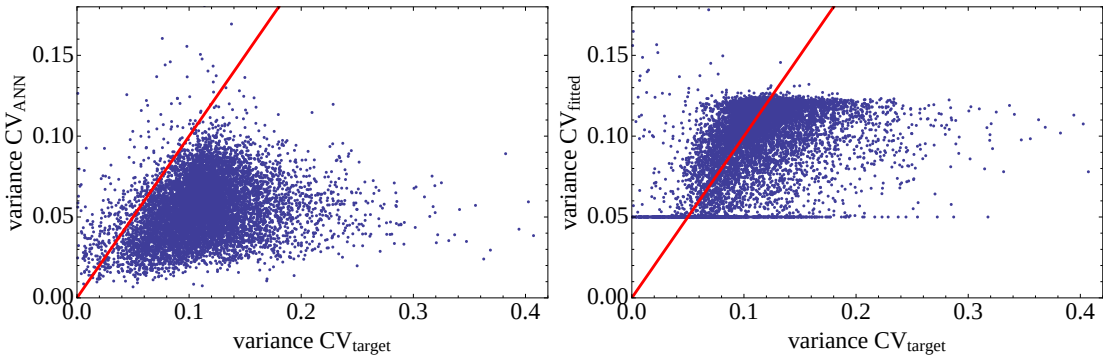


Figure 4.2: One the left hand side the variance of the profiles as output by the neural network are plotted over the variance of the target vectors, correlation coefficient is $r(\sigma_{\mathrm{CV}_{\mathrm{target}}}, \sigma_{\mathrm{CV}_{\mathrm{ANN}}}) = 0.235$. Correlation improves to $r(\sigma_{\mathrm{CV}_{\mathrm{target}}}, \sigma_{\mathrm{CV}_{\mathrm{fit}}}) = 0.385$ for the variance predicted using ansatz Eq. (4.1), as shown one the right hand side. The red lines mark the unity line of perfect prediction.

This predicted variance is stamped on the predicted CV output by the neural network. We call the resulting predicted contact vector predCV.
The scheme is quite generic and can be used to predict all kinds of profiles. The authors of the original paper e.g. apply it to predict secondary structure propensities. The predCV constitutes a dramatic improvement in prediction quality in comparison to the meanCV. The closer the prediction gets to the target CV the better the results of a sequence alignment, as discussed in the upcoming chapter.

5 Protein Sequence Comparison using predicted structural Profiles

The most common approach to perform pairwise sequence alignments consists of minimizing a penalty function that directly depends on the amino acids in the sequences and a fixed set of penalty values for opening and extending gaps. Doing so, the alignment of two amino acids is favoured the more similar they are. The values quantifying this amino acid similarity are empirically computed from sequence alignment databases like e.g. BLOCKS for the most commonly used BLOSUM matrices [Henikoff & Henikoff, 1992].
This kind of alignment is only meaningful for relatively closely related protein sequences, at about family level.[1] Too far beyond this level, sequences can be insignificantly correlated so that they cannot be distinguished from random pairs even though their respective structures might be very similar. This partly derives from the fact that substitution matrices implicitly assume single point mutations in the amino acid sequence while series of point mutations in the same site are common for longer evolutionary times. The choice of substitution matrix to apply, furthermore, depends on the evolutionary distance of the alignment to be performed, which might not be known beforehand.
As a result, not for all sequence pairs meaningful sequence alignments can be found and, even worse, the significance of a sequence alignment cannot be thoroughly assessed from the score of the alignment in all cases. That infers that even relatively high sequence similarity values might be meaningless from a structural point of view. In modern alignment algorithms this handicap is partly compensated by searching a sequence database to collect statistics for a given query sequence from which a tailor-made substitution matrix, a so-called Position Specific Scoring Matrix (PSSM), can be derived leading to much better results especially for remotely related sequences. The systematic pursued here is rather different. In this chapter the same alignment framework, including numerical parameter values, developed for structure alignments in Chapter 3 is used on the predicted structural profiles derived in Chapter 4 to perform sequence alignments.

[1] The notion of family level is defined differently for sequences and structures, here the structural definition according to SCOP is referred to.

This approach is fundamentally different as we aim to firstly predict the structural profiles from sequence data, to subsequently apply them to compute alignments. The prediction was achieved by training a neural network with the PSSMs output by PSI-BLAST [Altschul *et al.*, 1997]. The identical PSSMs will be used to compute PSI-BLAST's alignments for quality assessment, making the results perfectly comparable.
Stating that the quality reference for alignments for both structure and sequence derived profiles should depend on the structural superimposition, we use examples from structure databases and assess the alignment quality of our algorithm by the same structural measures used before to assess structure alignments. This makes sure that the sequence alignment results are *structurally relevant* and is much more objective than e.g. evaluating sequence alignments using structure alignments as done by Park *et al.* [1998] and others.
The results shown here are only partly comparable to other publications that are concerned with the assessment of alignment programs[2] in which usually either purely sequence derived scores are measured or sequence alignments are compared to structure alignments. The first approach suffers from the discussed insignificances of the scores, while the second intrinsically tests for the agreement with insecure references that might themselves be of limited quality.
Of course, following this path for the assessment does not set us free from defining a similarity significance score on pure sequence information, since coordinate data is not known in practical applications of sequence alignment tools.
An additional advantage of the approach to combine the predicted profiles with the generic alignment scheme is that alignment quality is expected to improve together with improved prediction quality, a prominent field of research today.

5.1 Protein Sequence Alignment using SABERTOOTH

The structure alignment framework developed in Chapter 3 is mostly generic relying only on some formal properties of the structural description. Besides of the vectorial form, applicable profiles need some kind of correlation with stability. But also representations based on other notions would be feasible as long as a cost function can be defined on them that can be optimized to identify good alignments.
Utilizing a prediction for the contact vector used for structure alignments before allows to keep the cost function together with the parameters just like applied there [Teichert *et al.*, 2009]. Only the post-processing step cannot be adopted since

[2] see e.g. Ahola *et al.* [2006], Lassmann & Sonnhammer [2002], Park *et al.* [1998]

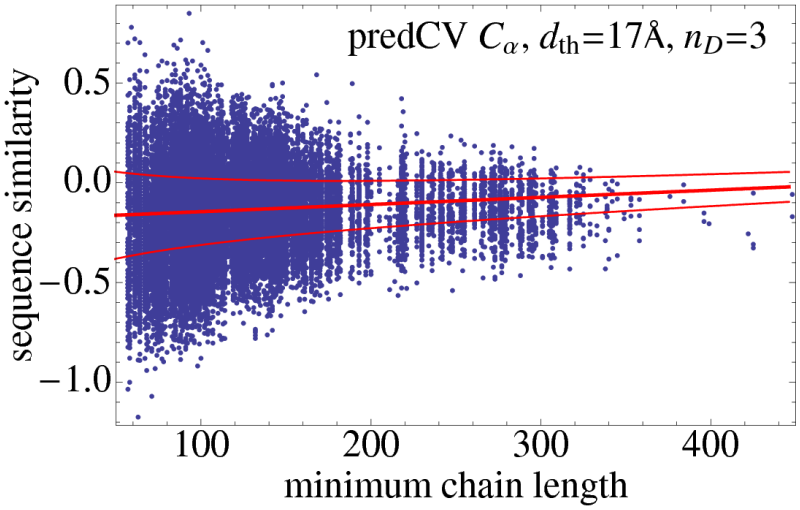


Figure 5.1: The scatter plot shows the seqSim over shorter chain length for the 31284 alignments of unrelated pairs in the Z-score set. The thick red line shows the $\langle \text{seqSim} \rangle$, the thin lines mark $\langle \text{seqSim} \rangle \pm \sigma_{\text{seqSim}}$.

it explicitly depends on the knowledge of coordinates. For the significance score, as well computed from the structural superimposition, a sequence based alternative is needed.

5.1.1 Defining a Significance Measure on Sequence Data

In perfect analogy to the definition of the structural and evolutionary Z-scores by analysing the random background in Section 3.2.2, a Z-score can also be formulated for sequence alignments. Of course, only the evolutionary Z-score on seqSim is applicable here since coordinate data is not at hand.
A power-law fit for mean $\langle . \rangle$ and standard deviation $\sigma_.$ of seqSim over the same set used before to fit the structural scores results to

$$\langle \text{seqSim} \rangle = -0.200101 \cdot \min(N_1, N_2)^{0.000359237} \tag{5.1}$$

$$\sigma_{\text{seqSim}} = 5.09614 \cdot \min(N_1, N_2)^{-0.678753} \,. \tag{5.2}$$

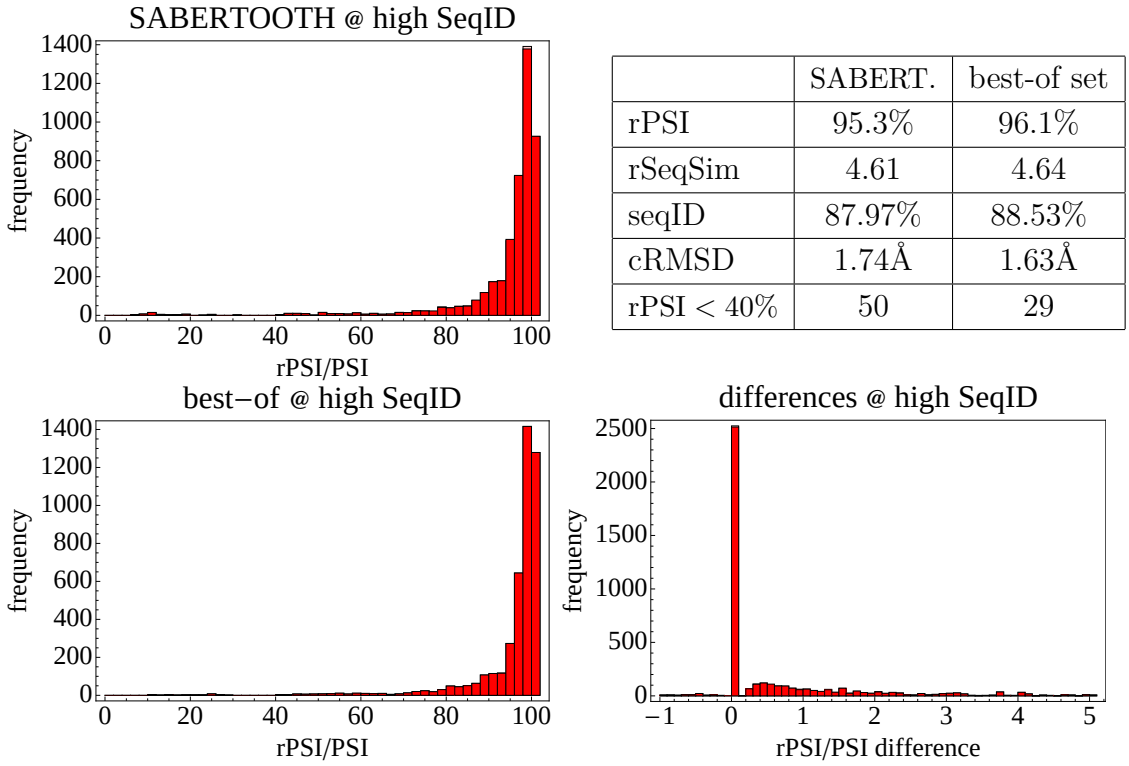
The resulting fits are depicted in Fig. 5.1. Comparing the fit values it shows that the mean seqSim expected from the background is lower and rises slower with chain length than for the structural equivalent. The standard deviation of the seqSim, in contrast, is substantially larger for the sequence alignment, as it could be expected.

5.2 Assessing Sequence Alignment Quality

The relevant measure to assess the quality of sequence alignments is not the sequence similarity produced by the alignment as an end in itself. The actual functional and evolutionary relationship is asked for but can only approximately be assessed if the structures are not know. In the test environment here, example proteins from a structure database can be used, hiding all structural information from the alignment tools (and also the profile prediction algorithm). The detailed coordinates re-enter the scene only for assessment.
Consequently, all test sets used in this chapter are exactly identical to those defined in Chapter 3, which has two advantages: The performance of the sequence alignment tools can directly be compared with the structure alignment tools. Furthermore, the comparison of the best-of references computed from sequence and from structure tools sheds light on the limitation of sequence alignment accuracy in general, when accepting that the structural best-of sets are quasi-perfect.
The sequence alignment tools used as references include ClustalW [Thompson *et al.*, 1994] and T-Coffee [Notredame *et al.*, 2000], very widely used standards in the field of pairwise alignments. Both utilize only information extracted directly from the sequences together with all-purpose substitution matrices. A more advanced technique is PSI-BLAST by Altschul *et al.* [1997]. Instead of direct pairwise alignments PSI-BLAST searches a sequence database for homologous sequences in order to compute a PSSM, a tailor-made substitution matrix. This PSSM is iteratively refined and finally applied to create the actual alignment. The computation of a PSSM is quite time consuming and can take up to some minutes depending on sequence length, database size, number of iterations, and other parameters.
The assessment of sequence alignments follows the same logic applied before to assess structure alignments with only one exception that demands an additional test: Alignments with very high sequence identity are the realm of sequence alignment tools, which is simply due to the fact that the alignment problem degenerates to a simple text search if sequences are largely identical. For the SABERTOOTH sequence alignment, in contrast, this cannot be taken for granted. The quality of the predicted contact vector introduces an additional source of possible difficulty. Although similar sequences are supposed to result to similar predicted profiles one has to make sure that the whole scheme works well in this obligatory regime.

5.2.1 Sequence Alignment at high Sequence Identities

To probe the performance of SABERTOOTH sequence alignments at very high sequence identities an additional test set was set up from a PDB clusters file with



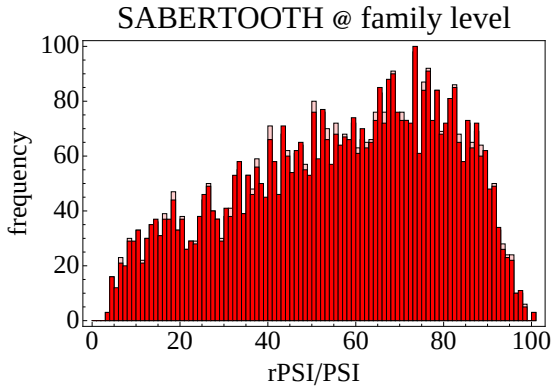
	SABERT.	best-of set
rPSI	95.3%	96.1%
rSeqSim	4.61	4.64
seqID	87.97%	88.53%
cRMSD	1.74Å	1.63Å
rPSI < 40%	50	29

Figure 5.2: All reference tools and also the best-of reference achieve these high values in rPSI, rSeqSim, and seqID. The difference histogram shows that 95% of all alignments are less than 5% points worst than the best-of reference.

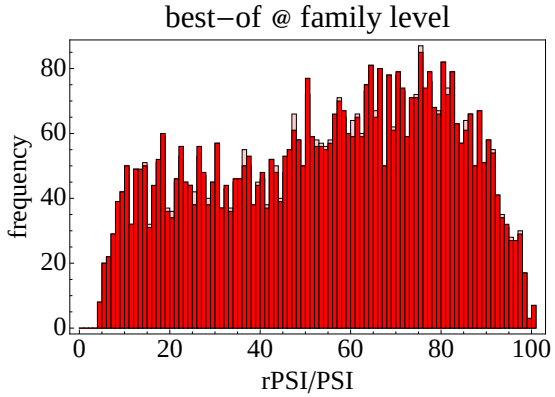
more than 50% pairwise sequence identities. From this set 100 clusters with more than 10 chains were selected randomly, abstracting short chains with less than 50 amino acids and also alignments of perfectly identical sequences. For all clusters the all vs. all combination of all members constitute the set of 4500 alignments, leaving 4408 after removing those for which not all reference tools output alignment results. Sequence identity just counts identical aligned amino acids, the result is normalized by the shorter sequence length. All considered reference tools achieve to assign about 88% of mean sequence identity and around 95% of mean rPSI, as already expected. The results for SABERTOOTH are roughly of same quality, as demonstrated in Fig. 5.2. The PSI difference histogram proves that 95% of SABERTOOTH's alignments are less than 5% points worse than the best-of reference, more than 2500 are identical. Fifty alignment examples out of the 4408 result in less than 40% PSI which is only 21 more than found for the best-of reference. High sequence identity test: Passed.

5.2.2 Similarity Recognition at different evolutionary Distances

Identical test sets and measures are used for the assessment of sequence alignments as were already used before for structure alignments, only the interpretation of the results is slightly different. Naturally much lower quality is expected from sequence alignments simply due to the reduced data input. This expectation is reflected also in the best-of reference, here computed from the established sequence alignment tools ClustalW, T-Coffee, and PSI-BLAST.
The sequence best-of reference assigns more than 20% points less rPSI in the mean on family and superfamily levels, and even 25% points less on the fold level. The error rate increases to about one third already on family level. From a structural point of view, sequence alignments on both superfamily and fold levels are close to random.
In contrast to that very large sequence similarities are found on superfamily level and still on the fold level slightly positive rSeqSim is assigned. This is no surprise since sequence similarity is the quantity maximized to compute the alignments. These high values are, nevertheless, contradictory to the results from the structure alignment test, where positive rSeqSim values were found only on the family level, even though much smaller. On superfamily and fold levels negative and strongly negative values were found, in accordance to the construction of the SCOP levels and coincidental with double sized values of rPSI in comparison to what is found here.
It is no question that the sequence alignment algorithms overestimate similarity by these artificially high values of rSeqSim. In fact, also the SABERTOOTH sequence alignment has this tendency. Values of sequence similarity are much higher than for the structure alignment but still it turns to SABERTOOTH's account here that it does not directly optimize mere sequence similarity. The rPSI values assigned are consistently higher for SABERTOOTH than for the best-of reference, about 2% points on family level and 5% on superfamily and fold levels, as shown in Figures 5.3 and 5.4.
The sequence similarity, even though slightly overestimated, nearly vanishes on the superfamily level and gets clearly negative on fold level.
The quality of the predicted profiles are responsible for that superior quality, together with the distinct ansatz. The maximum quality that can be reached in case the profiles could be predicted from the sequence perfectly can be assessed by applying the sequence alignment routine to the exact CV profiles instead of the prediction. Since the routines are identical, the results of this best-case test coincide with the structure alignment omitting the post-processing step. Doing so one finds that the mean rPSI values can be improved by about 5% points on family and superfamily levels and by even 7% points on fold level, while rSeqSim values are slightly decreased.



	SABERT.	best-of set
rPSI	54.1%	52.0%
rSeqSim	0.457	0.500
cRMSD	8.65Å	9.15Å
rPSI < 40%	1306	1514



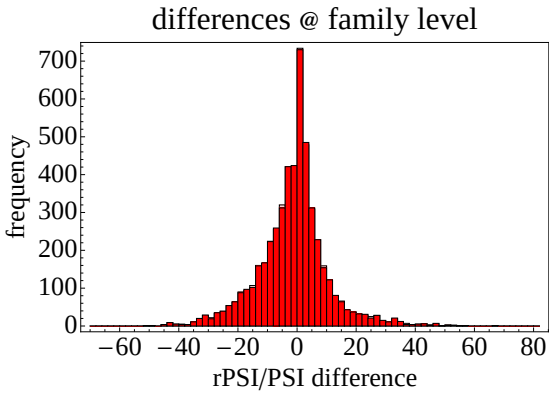


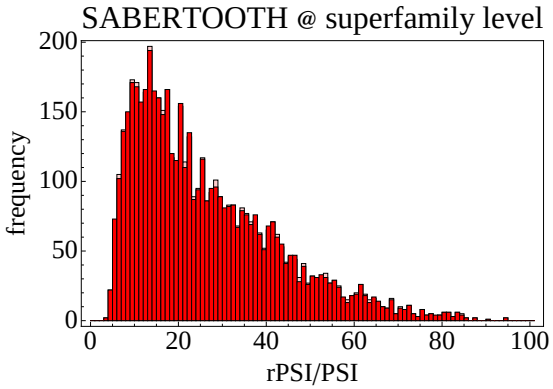
Figure 5.3: The rPSI distributions found over the family level test set are shown for SABERTOOTH and the best-of reference. The difference histogram proves that SABERTOOTH highly agrees with the reference in the similarities found and performs even better in many cases.

5.2.3 Comparison with established Sequence Alignment Tools

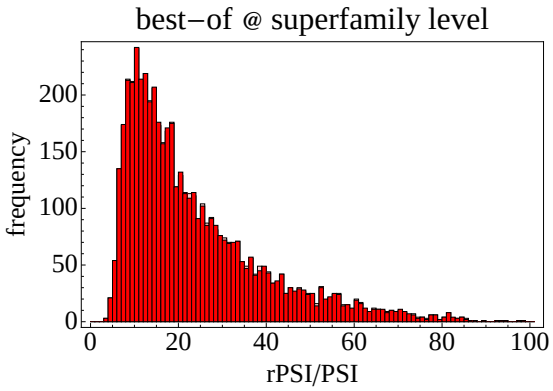
Like before, direct comparison is of interest to see how well single tools perform. While at the structural probing, DaliLite alignments were very close in quality to the best-of set, the picture is quite different for sequence alignments. All reference tools alone are far from the joint best-of quality. The full results are summarized in Table 5.1 on page 69.

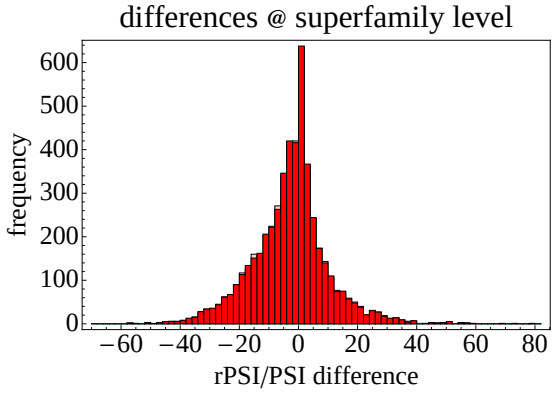
PSI-BLAST can take advantage from its superior ansatz only on the family level where it performs about 5% points better than ClustalW and T-Coffee. On superfamily level all three references reach very similar quality. Results on fold level are close to random, structural relations cannot be recognized anymore by sequence alignments whatsoever.

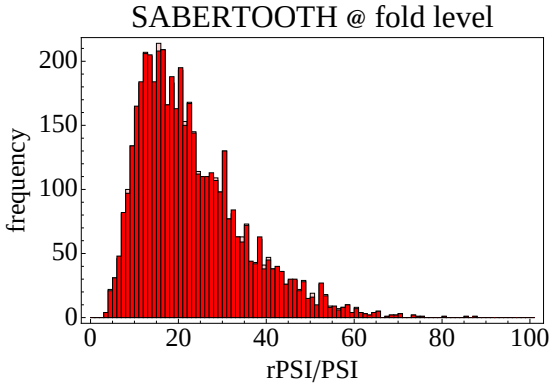
The low values of cRMSD found for PSI-BLAST are misleading. These values are result of the fact that PSI-BLAST does not align any sites that are outside the conserved core it identifies which reduces the percentage of aligned sites to 67%,



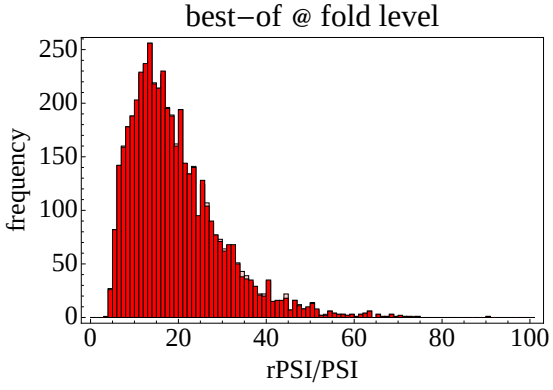
	SABERT.	best-of set
rPSI	24.0%	20.5%
rSeqSim	-0.006	0.102
cRMSD	16.70Å	18.11Å
rPSI < 30%	4005	4257







	SABERT.	best-of set
rPSI	20.3%	15.8%
rSeqSim	-0.121	0.027
cRMSD	15.75Å	19.75Å
rPSI < 20%	4273	4510



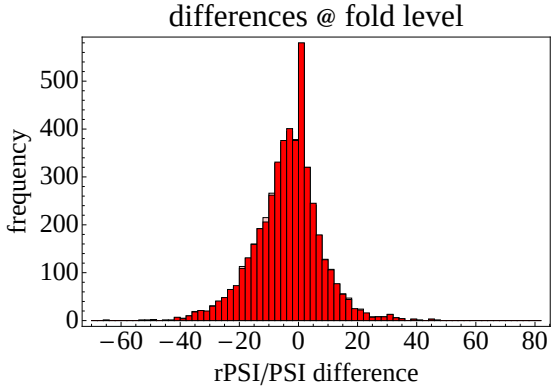


Figure 5.4: The rPSI distributions found over the superfamily and the fold level test sets are shown for SABERTOOTH and the best-of reference. SABERTOOTH performs significantly better measured in mean rPSI and error rate also for far homologues.

30%, and 20%, going from family to fold level and, in turn, reduces the cRMSD artificially. All other tools try to align the whole sequence resulting in percentages of aligned sites of 85–99%.
Another issue when using PSI-BLAST is that its alignment results are asymmetric in the choice of query sequence. It turned out that its performance can be improved when probing both orders and selecting the alignment that is assigned higher significance. All results shown for PSI-BLAST were computed that way.
The PSSMs that were input to the profile prediction scheme are identical to those used by PSI-BLAST to compute the alignments. The same parameters were applied to the same sequence database to compute them. Therefore, SABERTOOTH and PSI-BLAST are backed by the same data making the results shown here perfectly comparable.
SABERTOOTH based on predCV achieves consistently better results in all measures and over all test sets. This proves that making the detour to predict the contact vector and subsequently perform the alignment with parameters trained on the structure derived profiles was profitable. Especially placing emphasis on optimizing predicted structural instead of inherently sequence based properties can be hold responsible for the gain in structurally verifiable quality.

5.2.4 Structural Classification Abilities by Sequence Alignment

The classification abilities of PSI-BLAST are astonishingly on the background of its alignment quality. Although SABERTOOTH is nearly as sensitive, PSI-BLAST's coverage is much better on all three levels of similarity. The reason for this very likely derives from the much more sophisticated random sequence model that underlies its significance score. However, the advantage of PSI-BLAST in comparison to SABERTOOTH shrinks for far homologues. T-Coffee reaches reasonable classification quality only on family level, its score is close to random on superfamily and fold levels, as it could be expected from its rapidly degrading alignment quality. ClustalW performs well until about superfamily level taking into account that its data input is restricted to the two sequences without making use of a database search.
In general, classifications should only be derived from sequence alignments for family level similarities. Already the superfamily level is very demanding for all tools assessed here, as on display in Fig. 5.6 on page 68.

5.2.5 Computation Speed Comparison

Especially for sequence alignment tools, computation speed is a first order obligation since millions of sequences are known today. The same considerations apply for comparisons of speed as discussed for the structure alignment tools in Section 3.3.4. However, pairwise sequence alignments are usually much faster, since much less complicated representations and algorithms are used than for structure alignments. Sequences are of vectorial form and can be used as given, alignments are carried out simply by optimizing a scoring function with fixed substitution cost and gap penalties. Nevertheless, brute-force all vs. all alignment of a typical sequence database with some million sequences is infeasible.
A drawback of the SABERTOOTH approach is that the predicted profiles used for the alignments are very costly to compute. Depending on the size of the database used and the chain length of the query sequence, up to some minutes are needed only for the PSI-BLAST run to prepare the PSSMs. The subsequent neural network execution to get the predCV takes only little time compared to that. Although this has to be done only once per sequence and not twice per alignment, tremendous effort would be needed if a whole database of sequences has to be converted before the actual alignments can be processed.
The data shown in Fig. 5.5 include ClustalW, T-Coffee, and BLAST's tool *bl2seq* but not PSI-BLAST itself that was used for quality assessment. This is due to the

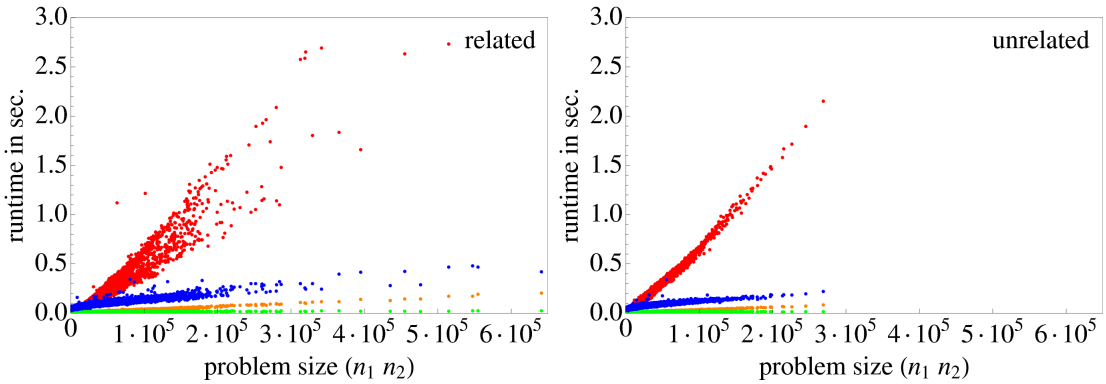
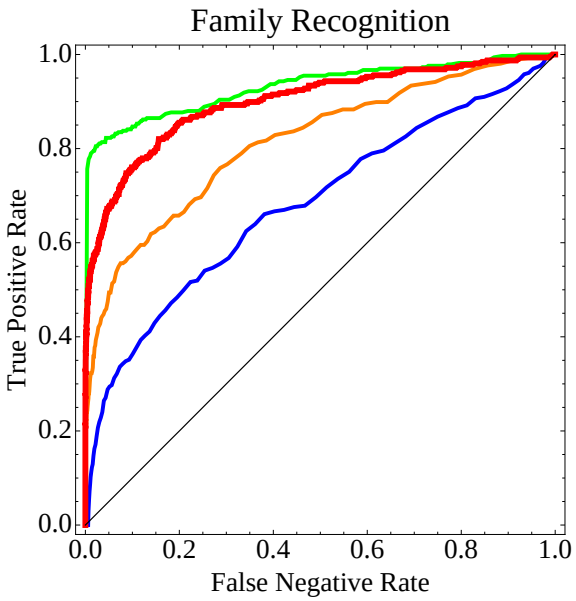


Figure 5.5: The scatter plots show runtimes over product of chain lengths. BLAST (green) is so fast that its CPU time consumption cannot be measure in milliseconds for most problem sizes. Also ClustalW (orange) and T-Coffee (blue) are very fast. The left plot refers to the family level test set used before containing related structures, the right plot refers to unrelated structures. SABERTOOTH is significantly slower than the other tools.

fact that PSI-BLAST is designed to align a query sequence to a database, even though the pairwise alignments and scores can be extracted, these runs are very tedious since a PSSM for the query sequence is computed first and is then aligned to the database. The high coverage in respect to far homologues is bought-in with some additional computational effort.
In contrast to that, bl2seq computes direct pairwise alignments of two sequences with the same algorithm used iteratively by PSI-BLAST and is, hence, more appropriate for this comparison of speed, even though its alignment quality is much worse. Consequently the runtimes shown for SABERTOOTH presume precomputed predCVs. The comparison of runtimes shows that SABERTOOTH is by far the slowest tool assessed here. The runtimes of BLAST are not even measurable for most problem sizes. Also T-Coffee, the slowest of the three references, takes less than half a second to align two sequences of about 500 amino acids in length. The longest SABERTOOTH alignment, on the contrary, takes nearly six seconds.



AUC in %	fa	sf	fl
SABERT.	90.00	76.89	65.77
ClustalW	81.41	67.62	56.68
PSI-BLAST	93.17	81.42	69.16
T-Coffee	68.13	53.15	47.84

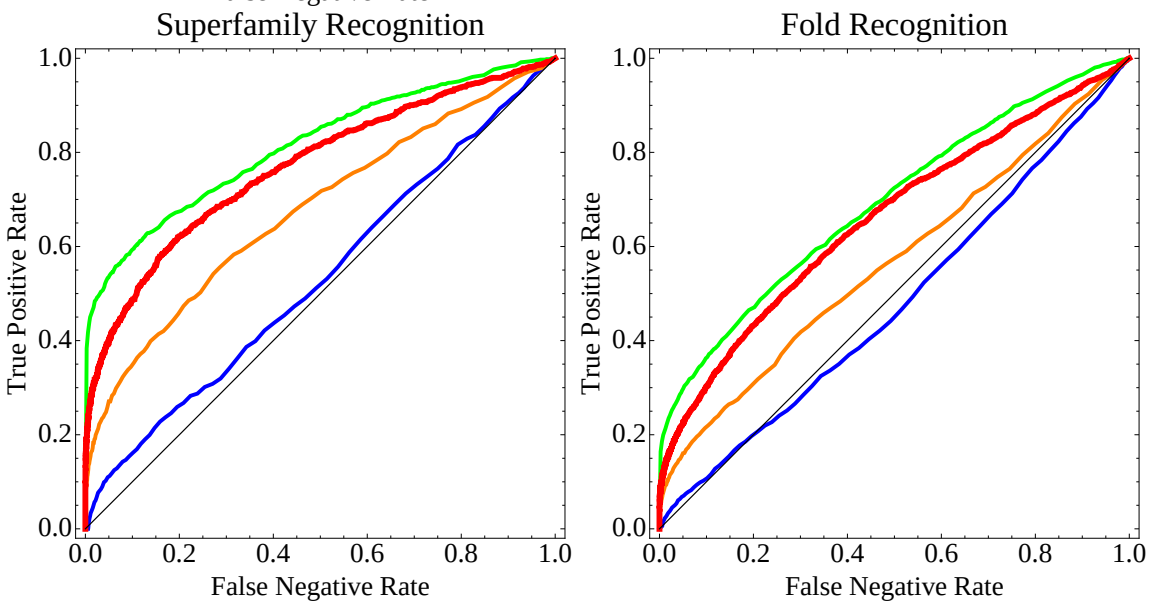


Figure 5.6: ROC-plots for the SCOP levels family (fa), superfamily (sf), and fold (fl) are shown. SABERTOOTH earns second place after PSI-BLAST that first of all achieves better coverage. The advance to SABERTOOTH shrinks for farther structural distances. T-Coffee is only applicable for family level recognition.

SCOP family level test set, 5014 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 40%
best-of ref.	52.0%	9.15Å	0.444	0.500	1514
SABERTOOTH	54.1%	8.65Å	0.475	0.457	1306
ClustalW	43.5%	13.70Å	0.397	0.553	2162
PSI-BLAST	48.3%	6.67Å	0.393	0.510	1831
T-Coffee	43.1%	13.52Å	0.386	0.365	2234

SCOP superfamily level test set, 4981 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 30%
best-of ref.	20.5%	18.11Å	0.207	0.102	3683
SABERTOOTH	24.0%	16.70Å	0.244	-0.006	3279
ClustalW	15.5%	23.47Å	0.201	0.218	4193
PSI-BLAST	15.3%	9.66Å	0.100	0.150	4162
T-Coffee	15.6%	23.50Å	0.195	-0.073	4239

SCOP fold level test set, 4737 alignments					
Program	rPSI	cRMSD	contOvr	rSeqSim	rPSI < 20%
best-of ref.	15.8%	19.75Å	0.175	0.027	2906
SABERTOOTH	20.3%	15.75Å	0.215	-0.121	2295
ClustalW	13.7%	22.64Å	0.176	0.173	3371
PSI-BLAST	8.9%	10.64Å	0.045	0.125	4337
T-Coffee	12.9%	23.30Å	0.164	-0.152	3600

Table 5.1: The table shows the detailed results for the assessment of alignment accuracy on the three test sets on different structural/evolutionary distances. rPSI refers to the relevant PSI suppressing small aligned fragments, cRMSD measures the mean square distance of all aligned sites, contact overlap is computed on heavy-atoms contact matrices with $d_{\text{th}} = 4$Å and $n_{\text{D}} = 3$. rSeqSim measures the relevant sequence similarity based on a BLOSUM62 matrix. In the last column the number of alignments below the respective rPSI cut-off is counted.

6 Improvement of structural Similarity Measures: FlexMaxSub

One of the most commonly used quantifiers for the structural similarity of protein pairs, the percentage of structural identity (PSI), has some intrinsic flaws. In some cases for pairs of actually nearly identical structures very low PSI values are assigned due to only minor distortions differentiating the two structures. From the mere PSI value, or a significance score derived from it, a pair might appear to be much less similar than it actually is. The reason for this misjudgement is to be searched for in the spatial superimposition that underlies not only the PSI but also the cRMSD and alike: Even when assuming a correct alignment, a single rigid-body rotation may not be sufficient to superimpose a pair of structures in the presence of internal movements.

Attempting to improve these measures one has to struggle with the definition of structural similarity. How to quantify the similarity of two structures that are identical besides of a single hinge tilting one of them, in comparison to indeed identical structures? Is there a difference in case a hinge moves only some secondary structure elements or a whole compact domain?

Just to allow for more than one independently rotated rigid-body does not solve the problem as desired if no strict rules have to be obeyed that define allowed movements. Obviously, just to allow up to N independent bodies in an N amino acid protein chain would trivially fold every structure onto every other, resulting in a meaningless PSI of 100% for every given pair. Consequently, some kind of compromise is asked for.

In this chapter the questions of above are boiled down to the definition of a *significant core* that is allowed to move independently from the rest of the structure in order to collect up similar fragments and minimize the overall cRMSD for cases that give meaningful contribution but no others.

From this multi-core superimposition standard measures for similarity can be computed, from which, in turn, significance scores can be derived that respect these movements improving their quality as classifiers.

In an alternative application, the newly defined scheme is applied to pairs of proteins known to be very similar beforehand to exemplify the method's ability to detect hinge regions in proteins.

6.1 The FlexMaxSub Scheme

The PSI is commonly computed from a single rigid-body rotation using the MaxSub algorithm by Siew *et al.* [2000] to define the relevant subset of aligned amino acids that form the common conserved core of the two structures in an alignment, as discussed in Section 3.2. But proteins are no rigid objects. Internal movements and distortions occur due to various reasons both in the individual life time of a protein as well as in the evolution of a family over a long period of time. Instantaneous changes in conformation happen on account of changed environmental conditions, e.g. changes in concentration of a ligand molecule or through the presence (or absence) of a binding partner. These influences often times result in vast structural movement. Evolutionary changes in the amino acid sequence, i.e. insertions, deletions, and amino acid substitutions, cause changes of all magnitudes. Very frequently, limited local distortions in a structure happen in response to slightly changed physiochemical properties of the chain after amino acid substitution. But also large deviations are possible that might even hinder the protein to fold properly at all.
In order to take these influences into account explicitly when computing a similarity measure, one would be demanded to, firstly, identify them and, secondly, define their relative impact on the specific measure in some intrinsic way. The first is a hard task, the second probably impossible, at least never objective.
These problems can partly be circumvented by devising a more heuristic ansatz, stating that changes in the shape of a protein structure can be partitioned into two different classes: In the course of evolutionary modification, a series of small changes in the sequence gradually changes a structure, the mapping from sequence to structural change is quasi-continuous. This class should naturally be accounted for in a similarity measure. In the traditional definition of the PSI this class of modifications is respected by setting the cut-off for close sites to 4Å, a quite generous distance that consciously disregards local deviations.
The second class includes conformational changes, large movements in structure without or only little modification in the protein's sequence or through cumulated evolutionary modification that results in a 'sudden' vigorous change in protein structure. An example for such a change could be a formerly complexed chain that looses its binding partner and settles down in a changed state of minimum folding free energy. In contrast to the first class, these changes are unique properties of the specimen that are not attainable by a continuous measure of structural similarity and should be tackled with tailor-made analyses.

At this point the definition of the significant core comes into play. The aim is to find independent marked-off parts of a structure that are self-contained objects in the sense that it is significant that their mutual similarity is not accidental but shows some inherent biological relationship.
If these parts are identified cautiously enough, all local movements from the first class as introduced above, are retained while those from the second class can be compensated by independent rigid-body rotations, and thereupon properly considered. Using the unchanged standard algorithm to compute the PSI on this newly defined superimposition leads to what we called the flexPSI.
In most cases, PSI and flexPSI are just identical if there is only one common core. But for some cases, only the first significant core contributing to the flexPSI is identical to the core building up the PSI, all further cores uncover formerly unrecognized similarities. In this sense, the PSI can be understood as the first order approximation of the newly defined flexPSI.
In the following sections the algorithm and parameters that define the notion of a significant core and in turn the flexPSI are discussed and assessed. The investigation was carried out in collaboration with Jonas Minning [Minning *et al.*, 2009].

6.1.1 Definition of the significant Core

The standard MaxSub algorithm by Siew *et al.* [2000] is designed to find the largest conserved common core in a given pairwise alignment and there is no need to alter it. However, a drawback of the algorithm gets vital when there is more than one such core. This second core is simply ignored by MaxSub, which results in assigning only little similarity to a pair of structures that are indeed very similar.
A straightforward way to deal with multiple cores is to iteratively apply the MaxSub routine as-is to that part of the alignment that is not yet part of a core. To avert a scattering of fragments, an additional condition on the cores found by MaxSub is imposed that restricts the number of continuously aligned sites that are farther away in space than a threshold, furthermore a minimum length should be demanded.
While iterating, MaxSub will find further cores as long as there are at least as many aligned sites left as used for seeding the search but not every such fragment identifies an independent core. Thus, the subtle point is where to stop the iteration, namely, which cores should be accepted and which rejected.
Here, the prosaic definition of the significant core from above can be implemented by imposing a condition on the preliminary cores identified by MaxSub that is based on a criterion of statistical significance. To make this decision the structural Z-score defined in Section 3.2.2 can be used as a good approximation. If the probed core comprises a whole domain, the exact statistics applied to the first core are also valid

for the second, since they are independent. For the case that the tested core is below domain level, e.g. a distorted motif, the statistics used underestimate the significance of the similarity found, because it is actually less likely to find significant similarity in the left-over part. That means that using the same Z-score function defined for the single rigid-body rotation also for the cores results in valuing similarity the stricter the smaller the core. However, this behaviour is very welcome when defining a most conservative measure.
The FlexMaxSub algorithm iterates the following scheme: A standard MaxSub superimposition is computed. The set of spatially close sites after this rotation is checked to comply with the requirements that it (a) consists of as many or more sites as specified in a minimum core length parameter $l_{\min}$ and that (b) the fragments in this core are continuous in the sense that not more that $p_{\max}$ adjacent sites are farther in space than $d_{\text{th}}^{\text{MaxSub}}$. All sites that obey these rules form the first preliminary core that is only accepted as a core if the Z-score computed independently for this part of the structure exceeds a threshold Z_{thr}. The iteration carries on assigning secondary cores with all sites that are not yet member of a core and terminates if no such core can be found anymore. Left-over sites that could not be assigned to any core are added to that core adjacent in sequence for which the total cRMSD is lower.

6.1.2 A FlexMaxSub based Significance Score

Once the multi-body rotations according to the rules above are performed, cRMSD and PSI can be computed on this special superimposition, which will in the following be called flexRMSD and flexPSI, respectively. Doing so is reasonable if the definition of the significant core was reasonable and only independently significant parts were rotated to just compensate major distortions.
With the same reasoning a Z-score can be computed from the flexPSI also using the statistics fitted for the standard PSI. It is again only approximately correct to do so. When the statistics was applied to the cores, similarity found was underestimated. In contrast to that, when applying it to the total flexible superimposition, similarity tends to be overestimated because additional similarity was induced that was not accounted for in the Z-score fitting procedure. There are two reasons why this tendency is acceptable. Firstly, even though overestimated, the result should be better than the strong underestimation in the single-body superimposition and, secondly, this is exactly what experts would do in such a case. Two structures that are very similar besides of a conformational change are assigned e.g. the same family in the SCOP classification.
Although it is highly approximate to apply the same Z-score function in all three

cases, on a single-core rotation, a preliminary secondary core, and the combined cores after multi-body rotation, this is the best that can be done. It seems impossible to gather reasonable statistics to re-fit a Z-score for a second, third core and so on considering that in the overwhelming majority of the cases only a single core is found.

6.1.3 Comparison with the original Definition

Devising a scenario for the statistical assessment of the gain of FlexMaxSub over MaxSub is not as unpretentious as it was for the classification test. Before classifying structures as in SCOP, chains are decomposed into domains with the result that most of the expected movements that could potentially be identified are suppressed. For whole protein structures, on the other hand, no classification is available.
Due to this and other reasons the flex$Z$-score cannot properly be assessed on the classification set: Firstly, only a very few cases are supposed to change because no domain movements are expected in a set consisting of mostly single-domain structures. Secondly, a positive hit can only result from an alignment between a distorted and an undistorted example of similar structures that was correctly assigned in the classification database. Furthermore, for exactly those examples correct alignments are needed, an additional source of error. Considering all these influences, the classification test cannot be used to fully assess the method's power.
In fact, when applying FlexMaxSub to the 123753 alignments of the classification test set nearly all alignments stay untouched. For a Z-score cut-off defining significant cores of $Z_{\text{thr}} = 2$ ($l_{\text{min}} = 10$ and $p_{\text{max}} = 5$) only 26 alignments are evaluated differently. While these changes have a positive net effect on the classification, their total impact is negligible.
In the majority of the cases already large Z-scores are assigned even increased flex$Z$-scores. In other cases modified loops connecting two helices are broken allowing to superimpose the helices.
The example ASTRALids d1k94a_ vs. d1tiza_ stands out in particular. The assigned PSI increases from 46.3% to a flexPSI of 91.0%, resulting in an increase in the significance from $Z = 1.81$ to flex$Z = 5.95$. This case is especially interesting since the high similarity was obviously overlooked in the SCOP classification, in which the two domains are assigned different families. That this appraisal is indeed wrong is supported by CATH [Cuff *et al.*, 2008, Orengo *et al.*, 1997] that assigns the two domains to the same homology cluster, the CATH analogue of a SCOP family. The example is shown in Fig. 6.1

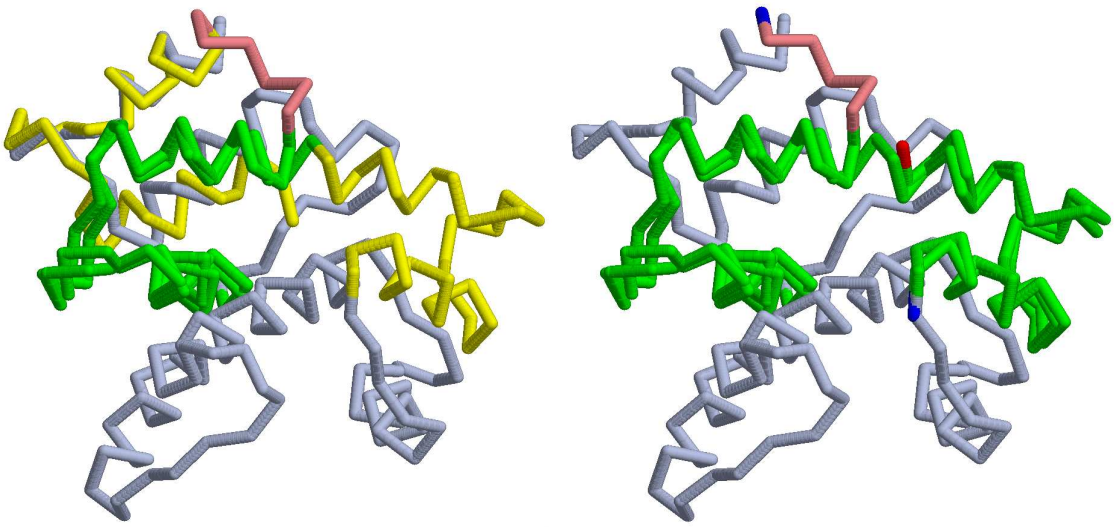


Figure 6.1: The figure shows the spatial superimposition of d1k94a_ vs. d1tiza_ applying a single-body rotation on the left hand side and a FlexMaxSub multi-body rotation on the right. Yellow backbone segments are aligned, green segments are close in space. In the FlexMaxSub superimposition all aligned pairs are also close in space which was achieved by rotating the secondary core about the red segment that marks a gap in the alignment. The alignment was computed using SABERTOOTH.

6.2 An alternative Application: Hinge Detection

From the point of view of rigid-body rotations FlexMaxSub searches for independent cores. From a different perspective the locations in which the cores are disconnected from each other are of interest by themselves. These locations are called hinges and identifying them is an active field of research. Several alignment programs predict hinges in order to improve their alignments. Examples are FATCAT by Ye & Godzik [2004] or RAPIDO by Mosca & Schneider [2008]. Before FlexProt [Shatsky *et al.*, 2004] computes an alignment, possible hinges are identified using HingeProt by the same authors [Emekli *et al.*, 2007].
What differentiates FlexMaxSub from, e.g. RAPIDO is the significance score imposed on secondary cores. This prevents to assign new cores until the total cRMSD falls under a threshold or alike and allows to run the algorithm against alignments of unknown similarity while RAPIDO is designed for very similar pairs only.
When run against such an example FlexMaxSub can be used to detect the hinge region. Figure 6.2 shows the flexible superimposition of the sequence identical chains PDBid 4clnA and 2bbmA ($N = 148$) in comparison to the rigid-body version. The

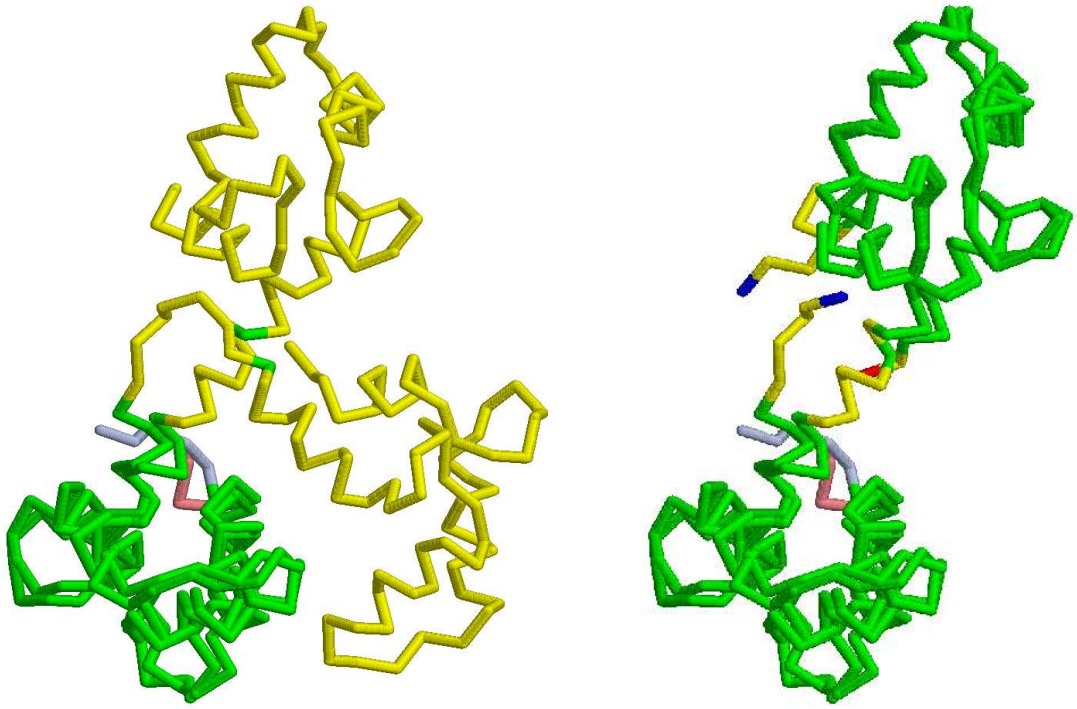


Figure 6.2: The figure shows the spatial superimposition of PDBids 4clnA and 2bbmA applying a single-body rotation on the left hand side and a FlexMaxSub multi-body rotation on the right. Yellow backbone segments are aligned, green segments are also close in space. The hinge region was correctly assigned between residues 79/80 in both chains.

SABERTOOTH alignment recognizes the structural deviation at the N-terminus of the structures and inserts a gap. The alignment results to a PSI = 47.3% leading to $Z = 6.10$ and the total cRMSD = 25.64Å. FlexMaxSub recognizes the hinge region between the two coils and cuts the structures between residues 79/80. The independent rotation reduces the total cRMSD to 2.52Å while the flexPSI reaches 90.4%, giving flex$Z = 14.30$.

7 Discussion and Conclusions

The intriguing interplay of tremendously complex protein structure and the rather simple way of storing the underlying information in form of vectorial DNA or protein sequences suggested that it should also be possible to define a vectorial description of protein structure. It could be shown here that connectivity based profiles are one way to describe protein structure in such a vectorial form that could be generalized to become applicable also to large structures with internal modularity. That these structural representations encode protein structure to a large extent and their strong correlation with sequence representations was shown before for small non-modular proteins. That the structural representations contain enough information to perform state-of-the-art structure alignments is a result of this thesis. Furthermore, it was shown that the correlation between sequence and structure can be exploited to predict structural representations from the sequence, in a quality that is sufficient to perform state-of-the-art sequence alignments. In fact, the sequence alignment performs even better than widely used standard tools as far as the aspect is on identifying structurally relevant similarities, an advantage bought in with higher time consumption in comparison to standard tools.
In general both, sequence and structure alignments, answer the same question about the evolutionary and functional relationship of a pair of proteins. Regardless of available data, either in form of structural coordinates or in form of the sequence, the question should also be addressed from this consolidated point of view. This perspective is also suggested by the coherence of sequence and structure.
The alignment scheme introduced here achieves this unification. For structure and for sequence alignments the identical algorithm is utilized together with the same set of parameters. As a matter of course, sequence alignments are less accurate also within this scheme since information is lost when predicting the structural profile from the sequence. Besides of that the profile alignment cannot be refined when the coordinates are not known.
The third flavour of alignments, namely the direct comparison of sequence and structure, that would round off the alignment tool presented is also feasible without further effort. However, the quality achieved is roughly identical to sequence to sequence alignments and cannot compete with current threading algorithms which is why the discussion of that point was skipped here. The gap in quality can poten-

tially be closed with an alignment post-processing step, in analogy to the structure alignment, that minimizes an energy functional of sequence and structure or through homology analyses.
Today, a large number of tools in particular for sequence but also for structure alignments exist while their quality, as referred to alignment accuracy and usability as a classifier, is not systematically assessed and compared, leaving the choice of a specific tool for a given task to a matter of taste. A large scale analysis of this, not at all trivial domain would help to improve results of all endeavours that use alignment programs as a tool. A first step to do so has been done in this thesis but definitive answers could only be given by joint efforts in a project dedicated to alignment quality, similar to e.g. CASP that is concerned with protein structure prediction quality. Of primary importance for this area is also a discussion about objective similarity measures, as touched here.
After all it can be concluded that the information about structure and function of a protein, that is stored in the correlation of the amino acid sequence, can be perceived in the contact network of the structure that in turn can be codified into vectorial profiles. These profiles, predicted or computed from the contact matrix, are sufficient to perform high quality alignments. Hence, the main goal to combine sequence and structure alignment within the same straightforward scheme was reached. The quality gap between sequence and structure alignments thereby quantifies our level of ignorance.

A Appendix

A.1 Amino Acid Residue Types

Residue Name	Descriptors		HP value	meanCV
Alanine	A	Ala	0.411633	1.03521
Cysteine	C	Cys	0.549505	1.14845
Asparagine	D	Asp	0.151707	0.887365
Glutamine	E	Glu	0.226615	0.863614
Phenylalanine	F	Phe	0.682647	1.11579
Glycine	G	Gly	0.228586	0.965902
Histidine	H	His	0.329858	1.00671
Isoleucine	I	Ile	0.692160	1.14399
Lysine	K	Lys	0.264855	0.864215
Leucine	L	Leu	0.700132	1.10246
Methionine	M	Met	0.449699	1.08215
Asparagine	N	Asn	0.240472	0.919536
Proline	P	Pro	0.276945	0.904669
Glutamine	Q	Gln	0.307464	0.915943
Arginine	R	Arg	0.311284	0.938612
Serine	S	Ser	0.231749	0.963186
Threonine	T	Thr	0.333868	0.996584
Valine	V	Val	0.683354	1.12779
Tryptophan	W	Trp	0.511161	1.08651
Tyrosine	Y	Tyr	0.591681	1.06892

Table A.1: The table lists the 20 amino acid residue types alongside with their respective hydrophobicity values as used to compute the HP (see: Bastolla *et al.* [2005]) and the values used to compute the meanCV of the C_α trace with $d_{\mathrm{th}} = 17$Å and $n_{\mathrm{D}} = 3$.

Bibliography

Ahola, V., Aittokallio, T., Vihinen, M., & Uusipaikka, E. (2006). A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, *7*, 484.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Bastolla, U., Ortíz, A. R., Porto, M., & Teichert, F. (2008). Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences. *Proteins*, *73*(4), 872–88.

Bastolla, U., Porto, M., Roman, H. E., & Vendruscolo, M. (2005). The principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*, *58*(1), 22–30.

Bastolla, U., Porto, M., Roman, H. E., & Vendruscolo, M. (2006). A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank. *BMC Evolutional Biology*, *6*, 43.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., & Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Canutescu, A., Shelenkov, A., & Dunbrack, R. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, *12*(9), 2001–2014.

Chandonia, J., Hon, G., Walker, N., Lo Conte, L., Koehl, P., Levitt, M., & Brenner, S. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Research*, *32*, 189–192.

Cuff, A., Sillitoe, I., Lewis, T., Redfern, O., Garratt, R., Thornton, J., & Orengo, C. (2008). The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271.

Eddy, S. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, *22*(8), 1035–1036.

Emekli, U., Schneidman-Duhovny, D., Wolfson, H., Nussinov, R., & Haliloglu, T. (2007). HingeProt: Automated prediction of hinges in protein structures. *Proteins*.

Henikoff, S., & Henikoff, J. (1992). Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences*, *89*(22), 10915–10919.

Holland, T., Veretnik, S., Shindyalov, I., & Bourne, P. (2006). Partitioning Protein Structures into Domains: Why Is it so Difficult? *J. Mol. Biol.*, *361*(3), 562–590.

Holm, L., & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, *16*(6), 566–567.

Holm, L., & Sander, C. (1993). Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.*, *233*, 123–123.

Holm, L., & Sander, C. (1994). Parser for protein folding units. *Proteins: Structure, Function, and Genetics*, *19*(3), 256–268.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, *292*(2), 195–202.

Jung, J., & Lee, B. (2000). Protein structure alignment using environmental profiles. *Protein Eng.*, *13*(8), 535–543.

Kabakçioglu, A., Kanter, I., Vendruscolo, M., & Domany, E. (2002). Statistical properties of contact vectors. *Physical Review E*, *65*(4), 41904.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*(5), 922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *34*(5), 827–828.

Kinjo, A., & Nishikawa, K. (2005). Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. *Biophysics*, *1*(0), 67–74.

Kinjo, A., & Nishikawa, K. (2006). CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinformatics*, *7*, 401.

Larson, S., Snow, C., Shirts, M., & Pande, V. (2002). Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology. *Computational Genomics*.

Lassmann, T., & Sonnhammer, E. (2002). Quality assessment of multiple alignment programs. *FEBS Letters*, *529*(1), 126–130.

Leo-Macias, A., Lopez-Romero, P., Lupyan, D., Zerbino, D., & Ortiz, A. R. (2005). An analysis of core deformations in protein superfamilies. *Biophys J*, *88*(2), 1291–1299.

Lesk, A. M. (2002). *Introduction to Bioinformatics*. Oxford University Press.

Lupyan, D., Leo-Macias, A., & Ortiz, A. R. (2005). A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, *21*(15), 3255–3263.

Minning, J., Teichert, F., Bastolla, U., & Porto, M. (2009). FlexMaxSub. (in preparation).

Mosca, R., & Schneider, T. (2008). RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic Acids Research*, *36*(Web Server issue), W42.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, *247*, 536–540.

Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*, *48*, 443–453.

Notredame, C., Higgins, D., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, *302*(1), 205–217.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., & Thornton, J. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1108.

Ortiz, A. R., Strauss, C. E., & Olmea, O. (2002). MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. *Protein Science*, *11*(11), 2606–2621.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, *284*(4), 1201–1210.

Porto, M., Bastolla, U., Roman, H. E., & Vendruscolo, M. (2004). Reconstruction of Protein Structures from a Vectorial Representation. *Phys. Rev. Lett.*, *92*(21), 218101.

Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, *20*(9), 374–376.

Shatsky, M., Nussinov, R., & Wolfson, H. (2004). FlexProt: Alignment of Flexible Protein Structures Without a Predefinition of Hinge Regions. *Journal of Computational Biology*, *11*(1), 83–106.

Shindyalov, I., & Bourne, P. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, *11*(9), 739–747.

Shirts, M., & Pande, V. (2006). Screen Savers of the World Unite! *Computing*, *10*, 43.

Siew, N., Elofsson, A., Rychlewski, L., & Fischer, D. (2000). MaxSub: An automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, *16*(9), 776–785.

Teichert, F., Bastolla, U., & Porto, M. (2007). SABERTOOTH: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, *8*, 425.
URL `http://www.fkp.tu-darmstadt.de/sabertooth/`

Teichert, F., Bastolla, U., & Porto, M. (2008). Protein Structure Alignment through a Contact Topology Profile using SABERTOOTH. *Proceedings of the German Conference on Bioinformatics GCB 2008, Lecture Notes in Informatics No. P-136, Gesellschaft für Informatik e.V.*, (pp. 75–84).

Teichert, F., Minning, J., Bastolla, U., & Porto, M. (2009). High Quality Protein Sequence Alignment combining Structural Profile Prediction and Structural Profile Alignment with SABERTOOTH. (in preparation).

Teichert, F., & Porto, M. (2006). Vectorial representation of single-and multi-domain protein folds. *The European Physical Journal B*, *54*(1), 131–136.

Thompson, J., Higgins, D., & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*, 4673–4673.

Tramontano, A. (2007). An account of the Seventh Meeting of the Worldwide Critical Assessment of Techniques for Protein Structure Prediction. *FEBS Journal*, *274*(7), 1651–1654.

Vendruscolo, M., Kussell, E., & Domany, E. (1997). Recovery of Protein Structure from Contact Maps. *Fold. & Des.*, *2*(5), 295–306.

Vullo, A., Walsh, I., & Pollastri, G. (2006). A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, *7*, 180.

Ye, Y., & Godzik, A. (2004). FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Research*, *32*(Web Server Issue), W582.

Zemla, A., Venclovas, Č., Moult, J., & Fidelis, K. (1999). Processing and analysis of CASP 3 protein structure predictions. *Proteins: Structure Function and Genetics*, *37*(s 3), 22–29.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.

Zhang, Y., & Skolnick, J. (2007). Scoring function for automated assessment of protein structure template quality. *Proteins*, *68*(4), 1020.

Résumé

Personal Data

Florian Teichert

Born 18.11.1975 in Frankfurt am Main, Germany
german, single

School Education

09/1982–07/1986	Karl-Nahrgang-Schule in Dreieich-Götzenhain, Grundschule des Kreis Offenbach/Main, Germany
09/1986–06/1995	Weibelfeld-Schule in Dreieich-Dreieichhain, Gesamtschule des Kreis Offenbach/Main

Civilian Service

03/1996–04/1997	Child care worker in the "Kindertagesstätte des DRK für behinderte Kinder Schloß Wolfsgarten" in Langen/Hessen, Germany

Academic Education

04/1997–11/2005	Studies of Physics at Technische Universität Darmstadt, Germany
11/2005	Diploma in Physics at Technische Universität Darmstadt, Germany
01/2006–03/2006 07/2007–08/2007 07/2008–07/2008	Research Stays at Universidad Autónoma de Madrid, Spain
06/2007	Conference 'Physical and Chemical Foundations of Bioinformatics Methods', Dresden, Germany

Professional Career

since 12/2005	Scientific employee at Technische Universität Darmstadt, Germany

List of Publications

Florian Teichert and Markus Porto. Vectorial representation of single-and multi-domain protein folds. *The European Physical Journal B*, 54(1):131–136, 2006.

Florian Teichert, Ugo Bastolla, and Markus Porto. SABERTOOTH: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, 8:425, 11 2007. URL `http://www.fkp.tu-darmstadt.de/sabertooth/`.

Ugo Bastolla, Angel R Ortíz, Markus Porto, and Florian Teichert. Effective connectivity profile: A structural representation that evidences the relationship between protein structures and sequences. *Proteins*, 73(4):872–88, 2008.

Florian Teichert, Ugo Bastolla, and Markus Porto. Protein Structure Alignment through a Contact Topology Profile using SABERTOOTH. *Proceedings of the German Conference on Bioinformatics GCB 2008, Lecture Notes in Informatics No. P-136, Gesellschaft für Informatik e.V.*, pages 75–84, 2008.

Jonas Minning, Florian Teichert, Ugo Bastolla, and Markus Porto. FlexMaxSub. 2009. (in preparation).

Florian Teichert, Jonas Minning, Ugo Bastolla, and Markus Porto. High Quality Protein Sequence Alignment combining Structural Profile Prediction and Structural Profile Alignment with SABERTOOTH. 2009. (in preparation).

Acknowledgements

First of all I want to say thank you to my doctoral adviser Markus Porto for scientific guidance, weekend availability, and much more encouragement than one can expect. Many thanks go to Ugo Bastolla for a lot of second opinions, scientific input, kind hospitality, and mountain trips. You have become a real friend. Thanks also to some of the nice people in your team: Alberto, David, and Raul in alphabetical order. I also want to thank Jonas Minning for the good collaboration over the last year.
Thanks a lot to my girlfriend Daniela Neumann for nourishing me, giving me shelter and loving support not only but especially in the last time.
The final thank you goes to my parents for an abundance of trust, backing and maybe also patience.

Eidesstattliche Erklärung

Hiermit erkläre ich eidesstattlich, daß ich die vorliegende Dissertation selbständig verfaßt, keine anderen als die angegebenen Hilfsmittel verwendet und bisher noch keinen Promotionsversuch unternommen habe.

Darmstadt, 2.12.2008